Międzynarodowy Instytut Biologii Molekularnej
i Komórkowej w Warszawie

Instytut Biochemii i Biofizyki
Polskiej Akademii Nauk w Warszawie

Łukasz Paweł Kozłowski

„Zintegrowany serwis bioinformatyczny do analizy białek. Przewidywanie domen i miejsc pozbawionych struktury trzeciorzędowej”

Rozprawa doktorska wykonana
w Laboratorium Bioinformatyki i Inżynierii Białka
w Międzynarodowym Instytucie Biologii
Molekularnej i Komórkowej w Warszawie

Promotor:
prof. dr hab. Janusz M. Bujnicki

Podziękowania

Pragnę podziękować mojemu promotorowi prof. dr. hab. Januszowi Bujnickiemu za kilkuletnią opiekę, cenne rady i inspiracje, jakich mi udzielił podczas prac nad projektami oraz za możliwość uczestniczenia w kursach i konferencjach naukowych.

Dziękuję również wszystkim koleżankom i kolegom z Międzynarodowego Instytutu Biologii Molekularnej i Komórkowej, za wspaniałą atmosferę pracy i cenne dyskusje.
Szczególne podziękowania należą się zwłaszcza wcześniejszym opiekunom metaserwera GeneSilico, Michałowi Kurowskiemu i Andrzejowi Kamińskiemu.

Pragnę także podziękować rodzinie za wsparcie, którego mi udziela na wybranej przeze mnie ścieżce kariery naukowej.

Na koniec, chciałem bardzo serdecznie podziękować mojej ukochanej żonie Kamili i mojemu synowi Bartkowi za wyrozumiałość, pomoc i troskę, zwłaszcza w dniach w których powstawała niniejsza rozprawa.

Ponadto pragnę podziękować za finansowanie z następujących źródeł:

- *grant promotorski (grant NN301 190139);*
- *grant ministerialny "Polish-Spanish Special Grant "Computer prediction and simulation of RNA tertiary structure formation" (HISZPANIA/152/2006), 2007-2010*
- *6th Framework Programme "Network of Excellence (NoE)" (LSHG-CT-2005-518238); 2008-2010*
- *EXGENOMES 7th Framework Programme "Exgenome Molecular Enzymes" (286556); 2011-2013*

Spis treści

Podziękowania 2

1. Streszczenie 7

2. Wykaz skrótów 9

3. Wstęp 13

3.1. Struktura białka 13

3.1.1. Struktura pierwszorzędowa 14

3.1.2. Struktura drugorzędowa 16

3.1.3. Struktura trzeciorzędowa 16

3.1.4. Struktura czwartorzędowa 18

3.1.5. Inne biologicznie istotne właściwości białek 19

3.1.5.1 Struktury transbłonowe 19

3.1.5.2 Struktury splecionych helis 20

3.1.5.3 Mostki dwusiarczkowe 22

3.1.5.4 Dostępność reszt aminokwasowych dla rozpuszczalnika 22

3.1.5.4 Regiony wewnętrznie nieuporządkowane (pozbawione stabilnej struktury trzeciorzędowej) 24

3.1.5.5 Oddziaływanie białek z DNA i białek z RNA 27

3.2. Przewidywanie właściwości biochemicznych białek w oparciu na sekwencjach 28

3.3. Meta-metody do przewidywania właściwości białek 30

3.4. Białka odpowiedzialne za modyfikację końca 3′ mRNA 32

4. Cel rozprawy 35

5. Materiały i metody 38

5.1. Sprzęt komputerowy 38

5.2. Bazy danych 39

5.3. Oprogramowanie 40
5.3.1. Oprogramowanie do wizualizacji danych 40
5.3.2. Wykorzystane metody uczenia maszynowego 41
5.3.3. Inne oprogramowanie i biblioteki pomocnicze 41
5.3.4. Programy i serwisy internetowe stanowiące część metaserwera do rozpoznawania zwoju 41
5.3.4.1. Programy do przewidywania struktury drugorzędowej 42
5.3.4.2. Programy do przewidywania domen 45
5.3.4.3. Programy do przewidywania regionów wewnętrznie nieuporządkowanych 47
5.3.4.4. Programy do przewidywania dostępności reszt aminokwasowych dla rozpuszczalnika 49
5.3.4.5. Programy do przewidywania helis transbłonowych 50
5.3.4.6. Programy do przewidywania struktur splecionych helis 52
5.3.4.7. Programy do przewidywania oddziaływania białek z DNA i białek z RNA 52
5.3.4.8. Programy do przewidywania mostków dwusiarczkowych 54
5.3.4.9. Programy do wykrywania zwoju białka 54
5.3.5. Programy i serwisy internetowe stanowiące część serwisu internetowego do przewidywania regionów wewnętrznie nieuporządkowanych 57
5.3.6. Programy stanowiące część serwisu internetowego do przewidywania domen białkowych 57
5.3.7. Programy wykorzystane w trakcie analizy kompleksu odpowiedzialnego za obróbkę końca 3′ mRNA 58
6. Wyniki 59
6.1. Metaserwer GeneSilico 59
6.1.1. Struktura serwisu 64
6.1.2. Interfejs użytkownika 65

6.1.2.1. Programy do rozpoznawania zwoju....69
6.1.2.2. Programy do analizy innych własności białek....71
6.1.3. Struktura baz danych....72
6.1.4. Skrypty zarządzające uruchamianiem programów składowych....73
6.2. Przewidywanie wewnętrznego nieuporządkowania struktury....75
6.2.1. MetaDisorder – meta-metoda oparta na innych programach do przewidywania regionów wewnętrznie nieuporządkowanych....75
6.2.1.1. Definicja wewnętrznego nieuporządkowania....76
6.2.1.2. Zbiór testowy....77
6.2.1.3. Miary jakości przewidywania użyte w czasie uczenia i testowania metod....78
6.2.1.4. Konstrukcja i testowanie metody....80
6.2.1.5. Skuteczność metody w eksperymencie CASP8....84
6.2.2. MetaDisorder3D – meta-metoda oparta na występowaniu przerw w przyrównaniach sekwencyjnych najbliższych homologów....85
6.2.3. MetaDisorderMD = MetaDisorder + MetaDisorder3D....88
6.2.4. Program MetaDisorderMD2 a miara S_{ww}....88
6.2.5. Wyniki programu MetaDisorder w czasie eksperymentu CASP9....89
6.3. Przewidywanie domen w białkach....90
6.3.1. Definicja domeny i jej granic....90
6.3.2. Zbiór testowy....91
6.3.3. Implementacja metody....93
6.3.4. Miary służące do oceny klasyfikatorów....93
6.3.5. Przewidywanie liczby domen....94
6.3.6. Przewidywanie granic domen....96
6.4. Analiza ludzkich białek odpowiedzialnych za modyfikację końca 3′ mRNA....98

6.4.1. Przewidywanie regionów wewnętrznie nieuporządkowanych białek odpowiedzialnych za modyfikację końca 3′ mRNA 101
6.4.2. Domeny białek odpowiedzialnych za modyfikację końca 3′ mRNA 105
7. Dyskusja 109
7.1. Przewidywania regionów wewnętrznie nieuporządkowanych 109
7.2. Przewidywanie domen w białkach.....112
7.3. Białka odpowiedzialne za modyfikację końca 3′ mRNA – biologiczny przykład zastosowania programów do przewidywania domen i regionów wewnętrznie nieuporządkowanych 114
7.3.1. Regiony wewnętrznie nieuporządkowane i ich znaczenie 114
7.3.2. Domeny strukturalne w białkach kompleksu odpowiedzialnego za modyfikację końca 3′ mRNA.....115
7.3.3. Model białka CPSF3 przykładem możliwości zastosowania modelowania homologicznego. 116
7.4. Metaserver GeneSilico jako przykład dużego projektu bioinformatycznego 116
7.5. Celowość wykorzystania meta-metod 119
7.6. Statystyki użytkowania serwisów 120
7.7. Implementacja metod, upublicznienie wyników badań 122
7.8. Przyszłe kierunki badań 122
8. Podsumowanie.....126
9. Bibliografia 128
10. Dorobek naukowy.....142

1. Streszczenie

Białka są jednym z najważniejszych składników komórki. Znajomość ich struktury na poziomie atomowym pozwala zrozumieć mechanizm działania i ich funkcje. Idealną sytuacją jest rozwiązanie struktury za pomocą technik takich krystalografia rentgenowska czy spektroskopia magnetycznego rezonansu jądrowego. Niestety bardzo często nie jest to możliwe. Z drugiej strony aktualnie znanych jest blisko 90 tysięcy (stan na luty 2012) struktur białkowych. Stanowią one bezcenne źródło wiedzy na ich temat. Już od czasów rozwiązania pierwszej struktury badacze zastanawiali się czy nie da się wykorzystać aktualnej wiedzy, aby przewidzieć strukturę lub przynajmniej pewne jej aspekty w oparciu na strukturach znanych białek. Wraz z pojawieniem się komputerów zaczęto tworzyć coraz dokładniejsze modele, które próbują tłumaczyć najróżniejsze zjawiska obserwowane w białkach. W ten sposób powstały programy, których zadaniem jest przewidzieć strukturę drugorzędową, dostępność reszt aminokwasowych dla rozpuszczalnika, wewnętrzne nieuporządkowanie oraz wiele innych cech białek.

W niniejszej rozprawie przedstawiłem zintegrowany serwis bioinformatyczny, który stanowi wygodną w użyciu platformę pozwalającą na uruchomienie ponad 100 aktualnie istniejących programów przewidujących najróżniejsze aspekty białek. Serwis pozwala na łatwe porównanie wyników poszczególnych programów, a następnie stworzenie wyniku będącego konsensusem, który reprezentuje najbardziej prawdopodobne przewidywanie. Dodatkową zaletą takiego podejścia jest to, że mając dostęp do kilku programów, które przewidują tą sama cechę można spróbować stworzyć tzw. meta-metodę, która przeprowadzi inteligentną syntezę wyników wzmacniając poprawne przewidywania i tłumiąc te fałszywe. W efekcie powstaje nowa metoda o jakości lepszej niż jej komponenty składowe. W ten sposób powstał program do przewidywania wewnętrznego nieuporządkowania także zaprezentowany w niniejszej rozprawie. Opierając się na wynikach metod przewidujących brak uporządkowania oraz inne aspekty białek stworzyłem program, który za pomocą metod uczenia maszynowego (wykorzystano algorytmy genetyczne i sztuczne sieci neuronowe) przewiduje tą cechę lepiej niż jakikolwiek aktualnie dostępny program. Zostało to potwierdzone w trakcie dwóch kolejnych edycji międzynarodowego eksperymentu CASP w czasie, którego dokonano niezależnych testów i porównano program do ponad 20 innych.

Kolejnym problemem poruszonym w rozprawie jest problem przewidywania domen białkowych. Jak powszechnie wiadomo większość białek eukariotycznych zbudowana jest z

modułów zwanych domenami. Wykazują one dużą autonomię zarówno pod względem strukturalnym w trakcie procesu zwijania jak i pod względem funkcji. Ponadto patrząc na problem podziału białek na domeny od strony technicznej należy zwrócić uwagę, że większość programów do przewidywania struktury trzeciorzędowej białek działa dobrze jedynie dla fragmentów reprezentujących pojedynczą domenę. Podobnie, często rozwiązania struktury białek technikami doświadczalnymi możliwe jest jedynie dla fragmentów białek reprezentujących domeny. Z tych powodów przewidywanie domen *in silico* jest niezwykle ważne. W niniejszej rozprawie zaprezentowałem program komputerowy o nazwie DomainSVM, który opierając się na maszynie wektorów nośnych przewiduje lokalizację domen. Przewidywanie to bierze pod uwagę cechy takie jak: entropia, hydrofobowość, przewidywana struktura drugorzędowa, dostępność reszt aminokwasowych dla rozpuszczalnika, występowanie regionów wewnętrznie nieuporządkowanych oraz obecność bliskich homologów w bazie domen białkowych CATH.

Należy podkreślić, że wszelkie programy bioinformatyczne powstają w konkretnym celu jakim jest pomoc biologowi w analizie danych doświadczalnych lub generowanie hipotez roboczych, które można później zweryfikować doświadczalnie. Dlatego zawsze poza sprawdzeniem statystycznej poprawności przewidywań powinno się również zweryfikować użyteczność tworzonych narzędzi. W tym celu korzystając z przedstawionych w rozprawie programów dokonano analizy ludzkich białek odpowiedzialnych za modyfikację końca 3′ mRNA. Białka te tworzą kompleks złożony z około 30 białek. Jego funkcją jest cięcie i poliadenylacja końca 3′ mRNA. W ramach tej analizy zbadano występowanie domen i regionów wewnętrznie nieuporządkowanych oraz zbudowano modele homologiczne białek kompleksu. Najważniejszym odkryciem wynikającym z tej części badań jest ustalenie, że regiony wewnętrznie nieuporządkowane są niezwykle istotne dla działania całego kompleksu. Ponad 80% białek tego kompleksu posiada przynajmniej jeden region wewnętrznie nieuporządkowany o długości powyżej 30 reszt aminokwasowych (średnia dla całego proteomu ludzkiego wynosi 35%).

2. Wykaz skrótów

Nomenklatura użyta w rozprawie opiera się na polskim wydaniu „Biochemii" Berga, Tymoczko i Stryera (PWN, Warszawa, 2005; przekład zbiorowy po redakcją Zofii Szweykowskiej-Kulińskiej i Artura Jarmołowskiego) oraz książce „Podstawy bioinformatyki" Xiong Jin (Wydawnictwa Uniwersytetu Warszawskiego, Warszawa, 2009; tłumaczenie: Bujnicki Janusz M., Kasprzak Joanna M., Figiel Małgorzata i inni). W przypadku dalszych wątpliwości źródłem była komunikacja ustna z prof. Januszem Bujnickim.

Ze względu na brak polskiej terminologii w niniejszej rozprawie sformułowania „nieustrukturalizowany" i „nieuporządkowany" traktowane są jako synonimy i używane są wymiennie.

Å – Ångström, jednostka długości równa 10^{-10} m, nie jest jednostką układu SI, 1 Å = 0,1 nm

ACC – ang. *accuracy* – dokładność

ANN – ang. *artificial neural network* – sztuczna sieć neuronowa, jedna z technik uczenia maszynowego, wzorowana na budowie układu nerwowego organizmów żywych

AUC – ang. *area under curve* – pole powierzchni pod krzywą ROC, odzwierciedla sumaryczną jakość klasyfiaktora w całym zakresie czułości i specyficzności

B-factor – czynnik temperaturowy

BLAST – ang. *Basic Local Aligment Search Tool* – powszechnie używany program do wyszukiwania podobnych sekwencji w bazach danych. Istnieje wiele wariantów tego programu, takich jak BLASTN, BLASTP i BLASTX przeznaczonych do wyszukiwania różnych typów sekwencji

CASP – ang. *Critical Assessment of techniques for protein Structure Prediction* – międzynarodowy eksperyment odbywający się co dwa lata mający na celu sprawdzenie możliwości programów i ludzi w przewidywaniu struktury białek

CATH – baza domen białkowych oparta na podobieństwie strukturalnym

CDD – ang. *Conserved Domain Database* – Baza Domen Konserwowanych reprezentowanych w formie macierzy PSSM

CF IIm – ang. *mammalian cleavage factor II* – ssaczy kompleks czynników cięcia typu II

CF Im – ang. *mammalian cleavage factor I* – ssaczy kompleks czynników cięcia typu I

CPSF – ang. *cleavage and polyadenylation specificity factor* – kompleks czynników specyficzności cięcia i poliadenylacji

CstF – ang. *cleavage stimulation factor* – kompleks czynników stymulacji cięcia

DALI – program do wyszukiwania podobieństwa strukturalnego lub/i baza domen białkowych oparta na jego wyniku

DISPROT – baza danych dotyczących białek wewnętrznie nieuporządkowanych

DNA – kwas deoksyrybonukleinowy

FN – ang. *false negative* – wynik fałszywie ujemny, błąd II rodzaju; przewidywanie braku występowania danej cechy w sytuacji, gdy ona występuje

FP – ang. *false positive* – wynik fałszywie dodatni, błąd I rodzaju; przewidywanie występowania danej cechy w sytuacji, gdy ona nie występuje

GNU – ang. *General Public License* – licencja wolnego i otwartego oprogramowania

GRAVY – ang. *grand average of hydropathy* – średnia hydrofobowość

HMM – ang. *hidden Markov models* – ukryte modele Markova, statystyczna metoda klasyfikacji sekwencji zdarzeń

IDEAL – ang. *Intrinsically Disordered proteins with Extensive Annotations and Literature* – baza danych dotyczących białek wewnętrznie nieuporządkowanych

InterPro – baza domen białkowych oparta na podobieństwie sekwencyjnym

IUP – ang. *intrinsically unstructured protein* – białko wewnętrznie nieuporządkowane, pozbawione zdolności do samodzielnego spontanicznego przyjmowania jednej stabilnej struktury trzeciorzędowej

IUR – ang. *intrinsically unstructured region* – wewnętrznie nieuporządkowany region białka

MCC – ang. Matthews *correlation coefficient* – współczynnika korelacji Matthews'a

MobiDB – baza danych białek wewnętrznie nieuporządkowanych

MR – ang. *molecular replacement* – technika podstawienia cząsteczkowego

MSA – ang. *multiple sequence alignment* – przyrównanie wielu sekwencji

NCBI – ang. *National Center for Biotechnology Information* – Narodowe Centrum Informacji Biotechnologicznej w USA

NMR – ang. *Nuclear Magnetic Resonance* – magnetyczny rezonans jądrowy

PDB – ang. *Protein Data Bank* – baza danych doświadczalnie rozwiązanych struktur cząsteczek biologicznych (głównie białek i ich kompleksów z różnymi cząsteczkami)

Pfam – ang. ***p****rotein* ***fam****ily database* – baza rodzin białkowych zdefiniowanych w oparciu na pokrewieństwie ewolucyjnym

PROSITE – baza domen białkowych oparta na podobieństwie sekwencyjnym

PSI-BLAST – ang. *position-specific iterative BLAST* – wersja programu BLAST, która używa strategii iteracyjnego przeszukiwania bazy danych poprzez tworzenie lokalnego przyrównania wielu sekwencji, a następnie przekształcanie go w macierz PSSM, która jest wykorzystywana do wyszukiwania kolejnych sekwencji dodawanych do przyrównania

PSSM – ang. *position-specific scoring matrix* – pozycyjnie specyficzna macierz oceniająca

RBF – ang. *radial basis function* – radialna funkcja bazowa, jedna z najczęściej wykorzystywanych funkcji jądrowych w maszynie wektorów nośnych

RNA – kwas rybonukleinowy

ROC – ang. *receiver operating characteristic* – krzywa ROC, analityczny sposób oceny poprawności klasyfikatora, zapewnia ona łączny opis jego czułości i specyficzności

RPS-BLAST – ang. *reverse position-specific BLAST* – wersja programu PSI-BLAST, w której pojedyncza sekwencja używana jest do przeszukiwania bazy danych zawierającej rodziny białek zapisane w postaci PSSM (czyli odwrotnie do PSI-BLAST, gdzie PSSM używana jest do przeszukiwania bazy danych pojedynczych sekwencji)

RRM – ang. *RNA recogniction motif* – domena rozpoznająca RNA

RSA – ang. *relative solvent accessibility* – relatywna dostępność reszty aminokwasowej dla rozpuszczalnika

SANS – ang. *small angle neutrons scattering* – małokątowe rozpraszanie neutronów

SAXS – ang. *small angle X-rays scattering* – małokątowe rozpraszanie promieni X

SCOP – ang. *Structural Classification of Proteins* – baza domen białkowych oparta na pokrewieństwie ewolucyjnym

SDS-PAGE – ang. *sodium dodecyl sulfate polyacrylamide gel electrophoresis* – elektroforeza w żelu poliakrylamidowym w obecności dodecylosiarczanu sodu

SMART – ang. *Simple Modular Architecture Research Tool* – baza domen białkowych oparta na podobieństwie sekwencyjnym

SVM – ang. *support vector machine* – maszyna wektorów nośnych, jedna z metod uczenia maszynowego

S_w – jedna z miar służąca do oceny poprawności klasyfikatora

TN – ang. *true negative* – wynik prawdziwie ujemny; przewidywanie braku występowania danej cechy zgadza się ze stanem rzeczywistym

TNR – ang. *true negative rate* – synonym czułości

TP – ang. *true positive* – wynik prawdziwie dodatni; przewidywanie występowania danej cechy zgadza się ze stanem rzeczywistym

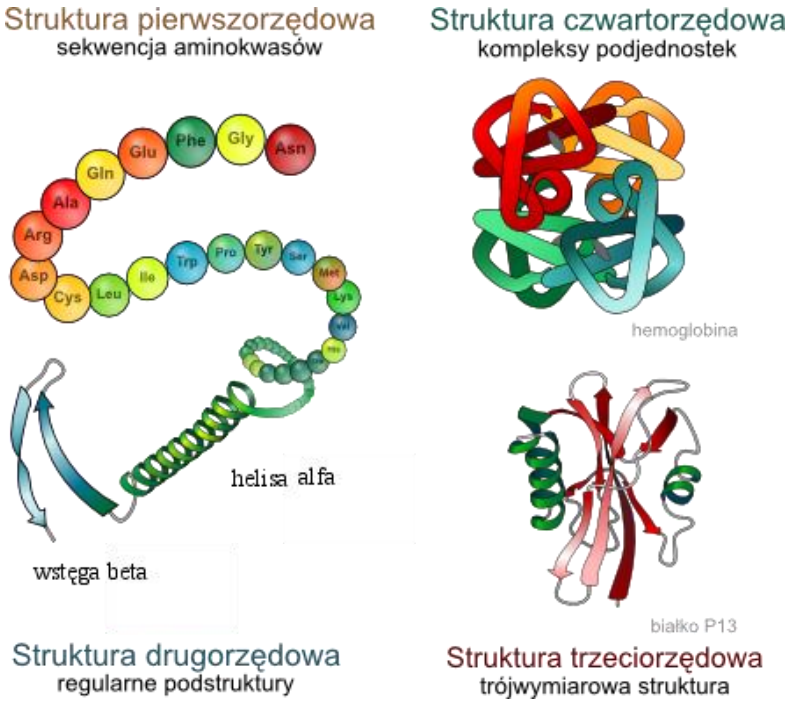
TPR – ang. *true positive rate* – synonym specyficzności

X-FEL – ang. *X-ray Free Electron Laser* – laser na swobodnych elektronach

3. Wstęp

3.1. Struktura białka

Białka stanowią główny składnik komórki i odpowiadają za większość czynności życiowych. Właściwie jedynym procesem, w który białka nie są bezpośrednio zaangażowane jest przechowywanie i przekazywanie informacji genetycznej, choć i w tym przypadku cały proces syntezy, transkrypcji i translacji przebiega w ścisłej zależności od białek takich jak polimerazy, czynniki transkrypcyjne i translacyjne. Pod względem chemicznym białka to wielkocząsteczkowe biopolimery zbudowane z reszt aminokwasowych połączonych ze sobą wiązaniami peptydowymi. Większość białek zbudowane jest z 20 aminokwasów; jedynie w niektórych białkach występują dodatkowe aminokwasy takie jak selenocysteina czy pirolizyna, ale ich występowanie jest ograniczone do określonych białek lub grup organizmów (Hao et al., 2004; Johansson et al., 2005). Obserwowany w komórkach przestrzenny kształt białka jest nieprzypadkowy i jest zdefiniowany przez sekwencję aminokwasową (tzw. strukturę pierwszorzędową) (Anfinsen, 1973). Decyduje ona o tym jak białko będzie się zwijać w struktury wyższego rzędu poprzez struktury drugorzędowe, aż po strukturę trzeciorzędową i czwartorzędową, które są jego końcową, biologicznie aktywną formą (ryc. 1).



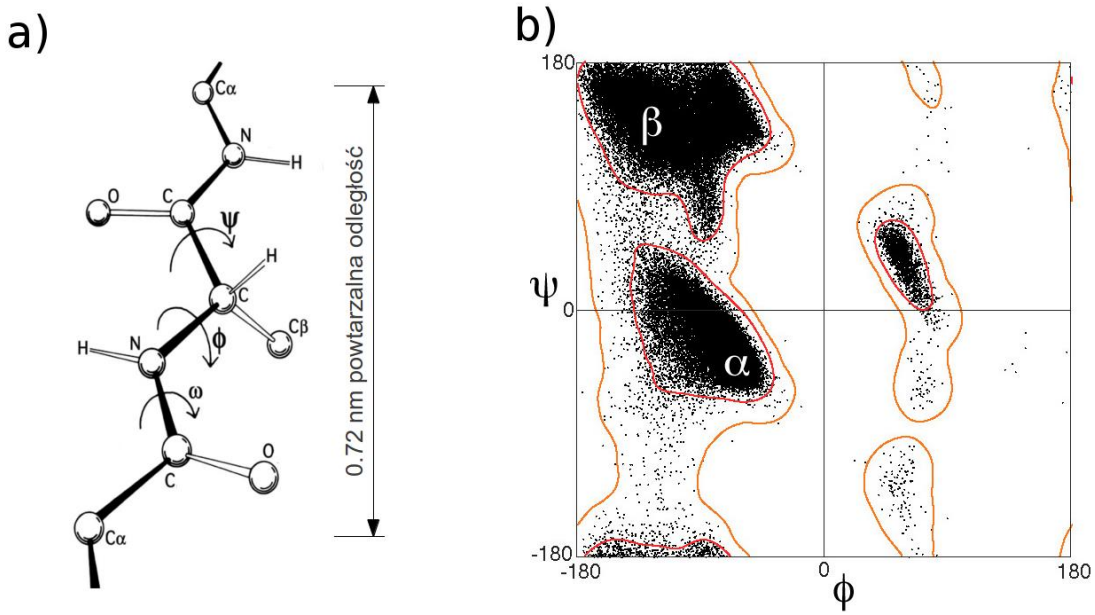
Ryc. 1. Poziomy złożoności struktury białek.

3.1.1. Struktura pierwszorzędowa

Sekwencja aminokwasowa białek nazywana jest strukturą pierwszorzędową i zapisujemy ją od końca N (reszta aminokwasowa z wolną grupą α-aminową) do końca C (reszta aminokwasowa z grupą α-karboksylową). Poszczególne aminokwasy łączą się ze sobą wiązaniem peptydowym (amidowym). Odległość między atomami węgla i azotu wynosi 0,133 nm. Ponadto wiązanie to wykazuje w znacznym stopniu właściwości wiązania podwójnego, co uniemożliwia swobodny obrót grupy >N–H w stosunku do grupy >C=O. Kąt obrotu (kąt torsyjny) wiązania N-H i C=O wokół wiązania peptydowego ≥C–N< oznacza się jako ω. Przyjmuje on wartości bliskie 180º lub (rzadko) 0º, co oznacza, że atomy obu ww. grup oraz atom węgla α związany z atomem azotu grupy aminowej leżą praktycznie w jednej płaszczyźnie. Jedynie między atomem węgla α a atomem azotu grupy aminowej oraz atomem węgla grupy karbonylowej (C′) występują pojedyncze wiązania, wokół których istnieje możliwość swobodnego obrotu. Odpowiednio, kąt torsyjny opisujący obrót grup atomów wokół wiązania $N–C^{\alpha}$ nazywamy kątem φ, natomiast kąt torsyjny opisujący obrót wokół wiązania $C^{\alpha}–C'$ - kątem ψ. Wartości przyjmowane przez kąty φ i ψ definiują możliwość przyjmowania przez łańcuch polipeptydowy struktur wyższego rzędu (ryc. 3.1.1) (Ramachandran et al., 1963).

Doświadczenia Anfinsena przeprowadzone na RNazie A doprowadziły do odkrycia, że struktura pierwszorzędowa decyduje o końcowym kształcie białka (Anfinsen et al., 1961). Należy zaznaczyć jednak, że prawidłowość ta dotyczy głównie małych, globularnych białek oraz znane są od niej wyjątki. Ciekawymi przykładami niejako przeczącymi temu zjawisku są białka, które przy tej samej sekwencji aminokwasów i w podobnych warunkach mogą przybierać różny kształt oraz białka, które do osiągnięcia funkcjonalnej formy wymagają dodatkowego udziału białek opiekuńczych (ang. *molecular chaperones*) (Ellis, 2006). Należy jednak podkreślić, że według hipotezy Anfinsena białko w warunkach fizjologicznych (określone zakresy parametrów fizykochemicznych takich jak pH, temperatura, czy też obecność innych cząsteczek w układzie) przyjmuje strukturę natywną, która odpowiada globalnemu minimum energii swobodnej całego systemu, czyli nie samego białka. Oznacza to, że przy zmianie warunków fizjologicznych (np. przy prowadzeniu do układu innej cząsteczki) środowisko się zmienia i struktura natywna także może się zmienić. To, czego hipoteza Anfinsena nie definiuje to sposób, w jaki stan natywny jest osiągany. Nawet w przypadku relatywnie krótkich białek o długości 100 reszt aminokwasowych liczba możliwych konformacji jest olbrzymia. W tym konkretnym przypadku przyjmując, że

mamy 99 wiązań peptydowych, które mają tylko trzy stabilne konformacje kątów φ i ψ, można obliczyć, że białko takie może mieć 3^{198} (czyli blisko $3*10^{94}$) konformacji przestrzennych. W związku z tym, jeśli proces zwijania białek polegałby na sprawdzaniu wszystkich możliwości i wyborze konformacji o najniższej energii, to nawet, jeśli przebiegałby on bardzo szybko (w ciągu nanosekund lub pikosekund na jeden wariant struktury) to czas ten byłby większy niż oszacowywany wiek wszechświata (Levinthal, 1968). Z obserwacji doświadczalnych wiadomo jednak, że białka zwijają się do struktury natywnej w ciągu milisekund, a czasem w ciągu mikrosekund i że łańcuch polipeptydowy wcale nie próbkuje „wszystkich możliwych" konfiguracji. Niestety mimo dużej ilości danych doświadczalnych i ogólnej znajomości praw fizyki odpowiedzialnych za proces zwijania białek, aktualny stan wiedzy nie pozwala przewidzieć szczegółów tego procesu *in silico*.



Ryc. 2. Fragment łańcucha peptydowego wraz z kątami. a) schemat obrazujący umiejscowienie kątów ω, φ, ψ w stosunku do wiązań w łańcuchu peptydowym; b) wykres Ramachandrana obrazujący występowanie określonych kątów torsyjnych dla 100 tys. reszt aminokwasowych z białek których struktury rozwiązano metodą dyfrakcji rentgenowskiej (rozdzielczość ≥ 1,8 Å i wartość czynnika temperaturowego ≤ 30, obramowanie wewnętrzne obejmuje obszar zawierający 98% danych, obramowanie zewnętrzne obejmuje 99,95% danych). Na wykresie zaznaczono regiony odpowiadające helisie α i wstędze β. Rysunek b na podstawie Proteopedia, Richardson JC, na licencji Creative Commons.

3.1.2. Struktura drugorzędowa

Polipeptydy białkowe często zwijają się do regularnych struktur typu helisy α oraz wstęgi β (Pauling i Corey, 1951). Helisa α jest skręconą strukturą, która stabilizowana jest przez wiązania wodorowe. Łańcuch główny polipeptydu tworzy cylindryczny rdzeń helisy α, natomiast łańcuchy boczne wystają na zewnątrz. Każda reszta aminokwasowa jest przesunięta w stosunku do sąsiedniej o 0,15 nm wzdłuż osi helisy i obrócona o kąt 100° wokół osi, w efekcie czego na jeden obrót helisy przypada ok. 3,6 reszt aminokwasowych i co czwarta reszta aminokwasowa w sekwencji jest położona blisko siebie w przestrzeni, a co druga reszta po przeciwnej stronie helisy. Ogólnie strukturę helisy α można zobrazować w formie prawoskrętnej śruby (helisy lewoskrętne są możliwe, jednak występują niezwykle rzadko, ze względu na obecność zawad sterycznych między łańcuchami bocznymi a szkieletem). Drugim powszechnym motywem struktury drugorzędowej jest wstęga β. Struktura β zdecydowanie różni się od helisy α: jest bardziej rozciągnięta, średnia odległość sąsiednich reszt aminokwasowych wzdłuż długiej osi jest większa i wynosi 0,35 nm. Łańcuchy boczne następujących po sobie reszt są zwrócone w przeciwnych kierunkach. Sąsiadujące ze sobą wstęgi β tworzą strukturę zwaną arkuszem β. W zależności od ułożenia poszczególnych wstęg możemy mieć do czynienia z równoległym i antyrównoległym arkuszem β. Arkusze zbudowane z antyrównoległych wstęg są stosunkowo płaskie, natomiast arkusze tworzone przez wstęgi równoległe są zwykle systematycznie skręcone.

Prócz wyżej wymienionych elementów struktury drugorzędowej, które zwykle stanowią największą cześć białka, zaobserwować można także inne, rzadziej występujące elementy takie jak helisy 3_{10}, helisy Π, pętle Ω i zwroty β (Kabsch i Sander, 1983). Generalnie struktury te pozwalają zmienić kierunek łańcucha polipeptydowego i są zlokalizowane na powierzchni białek, przez co często biorą udział w oddziaływaniach między białkami i innymi makrocząsteczkami.

3.1.3. Struktura trzeciorzędowa

Środowisko wodne, w którym białka się znajdują, definiuje dalszą ich organizację w struktury wyższego rzędu. Łańcuch polipeptydowy rozpuszczalnego w wodzie białka globularnego zwija się zazwyczaj w taki sposób, aby jego hydrofobowe łańcuchy boczne

znalazły się w rdzeniu białka, do którego dostęp wody jest ograniczony, natomiast łańcuchy polarne i naładowane zlokalizowane są na powierzchni białka. Z tego powodu wiele helis α i wstęg β ma charakter amfipatyczny, czyli reszty aminokwasowe jednej strony helisy lub wstęgi są w większości hydrofobowe i skierowane do wnętrza białka, a z drugiej strony reszty są hydrofilowe i skierowane do rozpuszczalnika otaczającego białko (Berg, 2007). Efektem końcowym są ściśle upakowane struktury globularne z niepolarnym rdzeniem. Określa się je mianem domen białkowych. Należy podkreślić, że niektóre białka zwijają się do struktur zbudowanych z dwóch i więcej domen, oddzielonych od siebie elastycznymi odcinkami o mniej lub bardziej zwartej strukturze. Pod tym względem białka można traktować jako makrocząsteczki o budowie modułowej, gdzie poszczególne domeny posiadają zazwyczaj pewną niezależność pod względem budowy wewnętrznej i samego procesu zwijania oraz funkcji biologicznej. Duża niezależność domen białkowych widoczna jest także na poziomie ewolucji, ponieważ białka o nowych funkcjach najczęściej powstają na drodze zmiany składu domen (np. poprzez rekombinacje fragmentów genów), ich duplikacji i zmiany funkcji poprzez nagromadzenie mutacji punktowych, a nie poprzez tworzenie nowego białka zupełnie „od zera" (ryc. 3.1.3.) (Ezkurdia i Tress, 2011). Podziału białek na domeny można dokonać uwzględniając wiele kryteriów i często jest on wysoce spekulatywny, dlatego też powstał szereg baz danych, które w mniejszym lub większym stopniu zgadzają się w adnotacji poszczególnych domen i lokalizacji ich granic. Bazy te uwzględniają różne cechy biologiczne i biochemiczne białek. Ogólnie bazy danych domen białkowych można podzielić na bazy strukturalne i sekwencyjne. Niektóre ważniejsze bazy strukturalne to:

CATH – http://www.cathdb.info/ – klasyfikacja domen opiera się głównie na oszacowywanym pokrewieństwie ewolucyjnym i związanym z nim podobieństwie sekwencji i struktur. Domeny wykrywane są na podstawie podobieństwa struktury drugorzędowej (przy wykorzystaniu programu CATHEDRAL (Pearl et al., 2003)) oraz struktury trzeciorzędowej (programy DETECTIVE (Swindells, 1995), DOMAK (Siddiqui i Barton, 1995)). W końcowym etapie część wyników jest sprawdzana ręcznie (Greene et al., 2007).

SCOP – ang. *Structural Classification of Proteins*, http://scop.mrc-lmb.cam.ac.uk/scop/ – baza danych w dużej mierze oparta na eksperckiej analizie pokrewieństwa ewolucyjnego i związanego z nim podobieństwa struktur oraz funkcji poszczególnych domen (Murzin et al., 1995).

DALI – http://ekhidna.biocenter.helsinki.fi/dali/start – baza danych oparta na podobieństwie struktur trzeciorzędowych wykryte za pomocą serwera DALI (Holm i Rosenstrom, 2010).
CDD – ang. *Conserved Domain Database*, http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtm – Baza Domen Konserwowanych, baza NCBI konserwowanych ewolucyjnie domen wykrytych za pomocą programu RPS-BLAST, domeny reprezentowane są w formie pozycyjnie specyficznych macierzy oceniających (PSSM) (Marchler-Bauer et al., 2011).

Niektóre ważniejsze bazy oparte na podobieństwie sekwencyjnym to:
Pfam – ang. ***p****rotein* ***fam****ily database*, http://pfam.sanger.ac.uk/ – białka dzielone są na rodziny potencjalnych homologów w zależności od przyrównania sekwencyjnego oraz podobieństwa ukrytych modeli Markova (ang. *hidden Markov models, HMM*). Także w tym przypadku niektóre etapy klasyfikacji przeprowadzane są ręcznie (Punta et al., 2012).
InterPro – http://www.ebi.ac.uk/interpro/ – baza zawiera informację na temat domen, rodzin białkowych oraz motywów sekwencyjnych, baza jest zintegrowana z wieloma innymi bazami m.in. z PFAM, CATH (Hunter et al., 2012).
PROSITE – http://prosite.expasy.org/ – regiony konserwowane ewolucyjnie wykrywane są za pomocą wzorców lub profili. Baza ta jest opiera się na automatycznym oznaczeniu domen pobieranych z bazy UniProtKB/Swiss-Prot (Magrane i Consortium, 2011). Ponadto baza zawiera adnotacje dotyczące miejsc aktywnych i mostków dwusiarczkowych (Sigrist et al., 2010).
SMART – http://smart.embl.de/ – zawiera ponad 1000 sprawdzonych przez ekspertów modeli domen, które można użyć do przeszukiwania bazy UniProtKB/Swiss-Prot oraz zsekwencjonowanych genomów (Letunic et al., 2012).

Przewidywanie domen białkowych na podstawie sekwencji aminokwasowej jest jednym z głównych celów niniejszej rozprawy doktorskiej.

3.1.4. Struktura czwartorzędowa

Najwyższy poziom tradycyjnej organizacji białka to struktura czwartorzędowa, opisująca oddziaływania pomiędzy niezależnymi łańcuchami polipeptydowymi tworzącymi funkcjonalną całość. Poszczególne łańcuchy nazywane są podjednostkami. Mogą one być identyczne (np. w homodimerze białka Cro bakteriofaga λ) albo mniej lub bardziej różnić się od siebie (dobrym

przykładem może być tutaj hemoglobina składająca się z dwóch podjednostek α i dwóch podjednostek β). W tym miejscu należy podkreślić, że przy pewnym poziomie złożoności systemów biologicznych zaczynamy mówić o kompleksach i systemach białkowych, które mogą składać się z bardzo dużej liczby heterogenicznych podjednostek (poza białkami w ich skład mogą wchodzić DNA, RNA i inne makrocząsteczki). Jako przykłady wielkich kompleksów/systemów makromolekularnych można wymienić np. rybosom, proteasom, kompleks poru jądrowego czy system składania transkryptów (ang. *spliceosom*). Skład i struktura takich kompleksów biologicznych może się zmieniać zarówno w czasie jak i przestrzeni (Berg, 2007; Griffin i Gerrard, 2012).

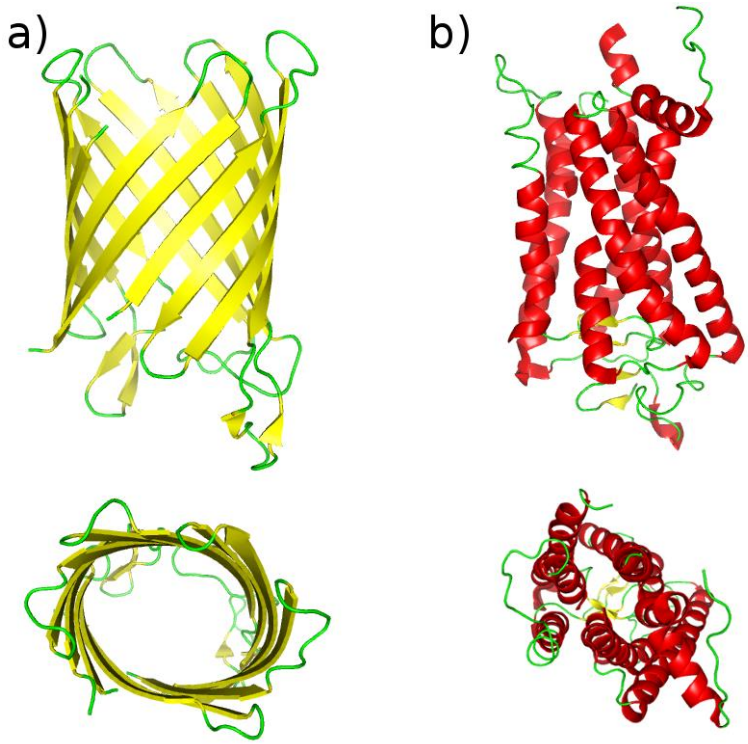
3.1.5. Inne biologicznie istotne właściwości białek

Oprócz wymienionych wyżej poziomów struktury białka istnieje szereg innych cech strukturalnych istotnych dla jego działania. Trudno je jednoznacznie przypisać do określonego poziomu lub są one specyficzną wariacją jednego z poziomów. Poniżej przedstawiono ważniejsze z nich.

3.1.5.1 Struktury transbłonowe

Charakterystyka białek lub domen występujących w błonach komórkowych jest zupełnie odmienna niż białek występujących w środowisku wodnym. Wynika to ze specyfiki otoczenia, w którym się one znajdują. W białkach transbłonowych mamy do czynienia z tendencją odwrotną w stosunku do tej obserwowanej u białek globularnych: reszty hydrofobowe skierowane są na zewnątrz i oddziałują z łańcuchami alkilowymi błony komórkowej, natomiast reszty polarne i naładowane skierowane są do wnętrza, tworząc w niektórych białkach transbłonowych kanał wypełniony wodą. Przykładem białek posiadających tego typu struktur są poryny, białka o zwóju nazywanym baryłką β, w całości zbudowanym z wstęg β zlokalizowanych w poprzek błony komórkowej (ryc. 3a). Innym często występującym typem białek błonowych są białka zbudowane jedynie z helis α. Przykładem takich białek jest rodopsyna, białko odpowiedzialne za przekazywanie sygnału świetlnego w narządach wzroku organizmów wyższych (ryc. 3b). Białka transbłonowe są szczególnie istotne, ponieważ działanie mniej więcej połowy obecnie używanych leków opiera się na blokowaniu ich działania. Ponadto, stanowią one około 27%

wszystkich białek kodowanych przez genom człowieka, jednak ze względu na swoje właściwości fizykochemiczne (m.in. uzależnienie struktury od oddziaływania z błonami biologicznymi) stanowią bardzo trudny cel dla badań doświadczalnych, o czym może świadczyć fakt, że jedynie 1% rozwiązanych doświadczalnie struktur w bazie danych PDB to struktury tego typu (Almen et al., 2009).

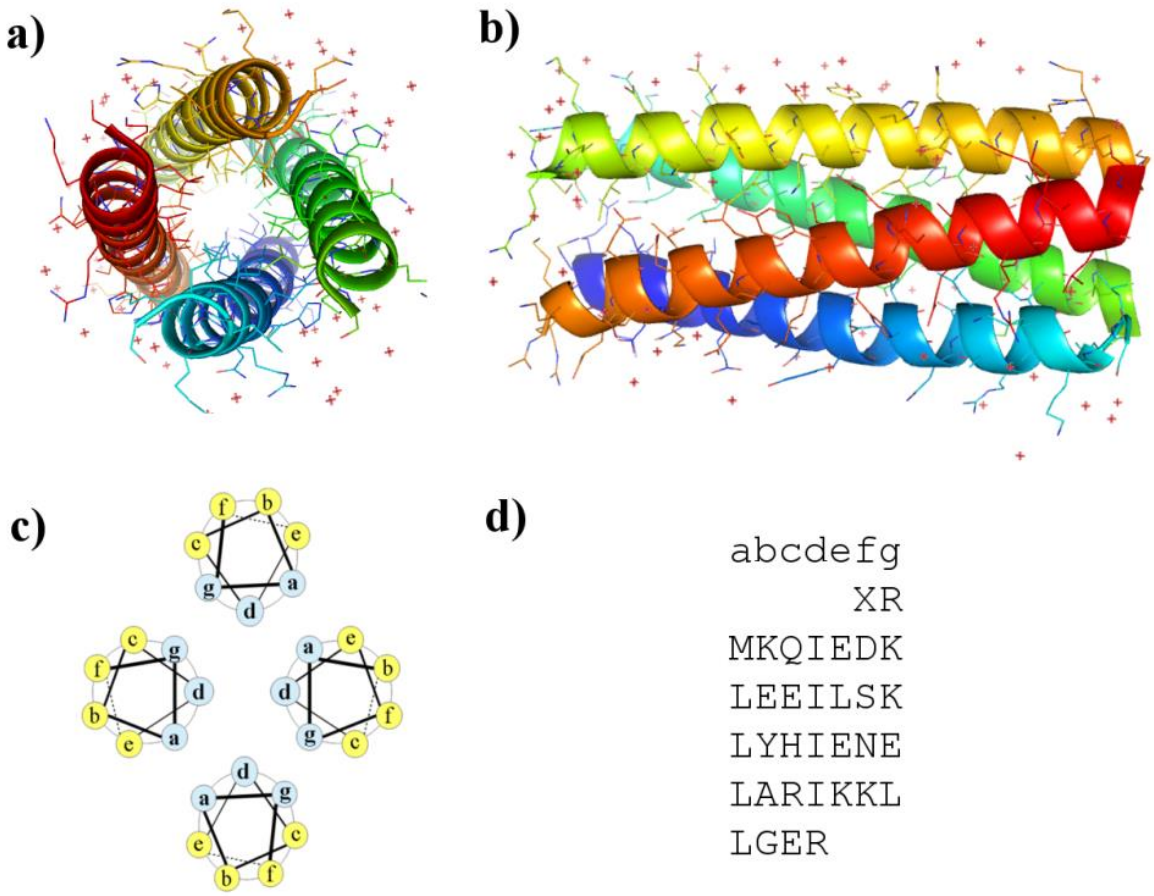


Ryc. 3. Typowe struktury białek błonowych. a) struktura poryny NanC, białka o zwoju baryłki β (pdb: 2WJR); b) struktura rodopsyny zbudowanej z helis α (pdb: 2I37). W panelu górnym przedstawiono widok w poprzek błony komórkowej, w panelu dolnym przedstawiono rzut z góry. Kolorem żółtym zaznaczono helisy α, kolorem czerwonym wstęgi β, pozostałe elementy struktury drugorzędowej zaznaczono na zielono. Rycinę wygenerowano za pomocą programu PyMOL (DeLano, 2002).

3.1.5.2 Struktury splecionych helis

Struktura splecionych helis, czasem nazywana superhelisą, to zbiór od dwóch do siedmiu helis α, które są splecione wokół siebie niczym nici w sznurze (ryc. 4). Wiele białek zawierających motyw splecionych helis zaangażowana jest w biologicznie ważne procesy takie jak regulacja ekspresji genów. Przykładami są białka onkogenne *c-fos*, *jun* oraz kolagen czy występująca w mięśniach tropomiozyna. Struktury splecionych helis zwykle zawierają wielokrotnie powtórzony motyw sekwencyjny *hxxhcxc*, gdzie *h* oznacza resztę aminokwasu

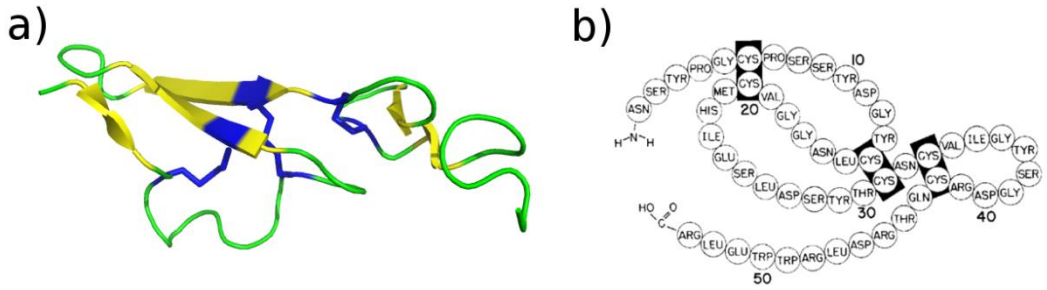
hydrofobowego, a *c* oznacza resztę naładowaną. Kolejne pozycje w obrębie tego motywu oznaczane są jako *abcdefg*. Miejsca *a-d* zajmują najczęściej izoleucyna, leucyna i walina. Sekwencja polipeptydowa zawierająca takie motywy ma tendencję do zwijania się w specyficzny sposób, ponieważ rozmieszczenie poszczególnych reszt aminokwasowych narzuca amfipatyczny charakter helis, które skręcają się prawie zawsze w lewo. W środowisku wodnym dla struktur tego typu najkorzystniejsze pod względem energetycznym jest zwinięcie się i oligomeryzacja w taki sposób, aby schować do wnętrza reszty hydrofobowe sąsiadujących ze sobą helis α. Ilość reszt hydrofobowych jest tak duża, że efektem końcowym są struktury mocno upakowane, niemalże w całości wysycone przez oddziaływania van der Waalsa (Grigoryan i Keating, 2008).



Ryc. 4. Struktura splecionych helis na przykładzie struktury zamka leucynowego białka GCN4 (pdb: 1GCL) a) widok z góry; b) widok z boku; c) topologia helis; d) motyw sekwencyjny *abcdefg* nałożony na sekwencję białka GCN4.

3.1.5.3 Mostki dwusiarczkowe

Jednym z bardziej specyficznych, aczkolwiek istotnych dla ostatecznej struktury wielu białek, wiązań jest wiązanie kowalencyjne między dwiema resztami cysteiny (energia wiązania wynosi około 60 kcal/mol). Oksydacja dwóch grup tiolowych pochodzących z tych reszt powoduje utworzenie tzw. mostka dwusiarczkowego. Należy podkreślić, że wiązanie to tworzy się najczęściej między resztami cysteiny, które w sekwencji są od siebie oddalone. Mostki dwusiarczkowe, o ile występują, są jednym z głównych czynników wpływających na końcową strukturę trzeciorzędowa oraz aktywność białka. Z jednej strony stabilizują odziaływanie między odległymi sekwencyjnie fragmentami białka, zaś z drugiej strony poprzez swój hydrofobowy charakter mogą stanowić ośrodek, wokół którego zaczną skupiać się inne reszty hydrofobowe w czasie zwijania się białka (ryc. 5). Liczba możliwych kombinacji wiązań dwusiarczkowych wzrasta wykładniczo wraz ze wzrostem liczby cystein w sekwencji białka. Przykładowo dla białka posiadającego 8 cystein ilość kombinacji wynosi 105 i zwykle jedynie jedna z nich występuje w biologicznie aktywnym białku (Wedemeyer et al., 2000).

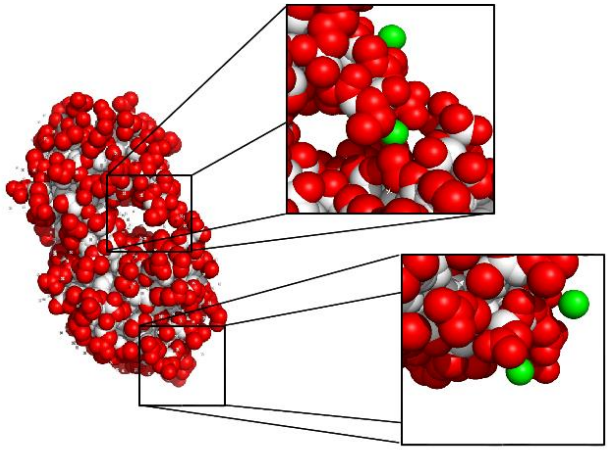


Ryc. 5. Czynnik wzrostu naskórka (EGF) jako przykład białka zawierającego mostki dwusiarczkowe. a) struktura przestrzenna mysiego EGF (pdb: 1EPH), kolorem niebieskim zaznaczono mostki dwusiarczkowe między resztami cysteiny; b) schematyczne położenie mostków dwusiarczkowych w sekwencji białka EGF, rycina sporządzona na podstawie (Savage et al., 1973).

3.1.5.4 Dostępność reszt aminokwasowych dla rozpuszczalnika

Jak wspomniano wyżej, ze względu na budowę grup bocznych reszty aminokwasowe różnią się od siebie pod względem hydrofilności, czyli tendencji do wiązania wody. Właściwość ta ma duży wpływ na strukturę drugorzędową oraz struktury wyższego rzędu białka, ponieważ definiuje ona preferowaną lokalizację reszt aminokwasowych w środowisku. Przykładowo: większość białek znajduje się w środowisku wodnym w związku z czym białko zwija się tak, aby reszty hydrofilowe były na zewnątrz. Analizując problem dostępności reszty aminokwasowej dla

rozpuszczalnika bardziej szczegółowo, łatwo możemy dojść do wniosku, że dostępna powierzchnia z jednej strony zależy od wielkości grupy bocznej (jej objętości), zaś z drugiej strony ograniczona jest przez właściwości sąsiadujących reszt. W związku z tym sformułowano pojęcie tzw. relatywnej dostępności reszty aminokwasowej dla rozpuszczalnika ang. *relative solvent accessibility* (RSA), która oznacza stosunek powierzchni tej reszty dostępnej dla rozpuszczalnika obserwowanej w strukturze danego białka do możliwej powierzchni maksymalnej wyznaczonej dla tripeptydu (Gly-X-Gly) – tabela 1. Reszty aminokwasowe uważa się za niedostępne dla rozpuszczalnika (zagrzebane), jeśli wartość RSA jest mniejsza niż określony próg (najczęściej stosowane progi to 0%, 5% i 25%). Wartość powierzchni dostępnej dla rozpuszczalnika wyznacza się stosując modelowanie komputerowe, które polega na „toczeniu" kulki odpowiadającej wielkością cząsteczce wody (1,4 Å) po powierzchni białka – ryc. 6.



Ryc. 6. Lizozym (pdb: 123L) przedstawiony w reprezentacji kulkowej z zaznaczonymi resztami dostępnymi dla rozpuszczalnika (kolor czerwony). Na zielono zaznaczono przykładowe cząsteczki wody, które toczone po powierzchni białka wyznaczają reszty dostępne dla rozpuszczalnika.

Tabela 1. Objętość i powierzchnia reszty aminokwasowej dostępna dla rozpuszczalnika dla tripeptydu Gly-X-Gly. Zestawienie danych z (Zamyatnin, 1972) i (Chothia, 1976).

Aminokwas	**Objętość**	**Powierzchnia**	**Aminokwas**	**Objętość**	**Powierzchnia**
A	88,6	115	L	166,7	170
R	173,4	225	K	168,6	200
D	111,1	150	M	162,9	185
N	114,1	160	F	189,9	210
C	108,5	135	P	112,7	145
E	138,4	190	S	89,0	115
Q	143,8	180	T	116,1	140
G	60,1	75	W	227,8	255
H	153,2	195	Y	193,6	230
I	166,7	175	V	140,0	155

3.1.5.4 Regiony wewnętrznie nieuporządkowane (pozbawione stabilnej struktury trzeciorzędowej)

Poza typowymi strukturami drugorzędowymi, które narzucają organizację struktur wyższego rzędu niektóre fragmenty białek, a czasem całe białka mogą być na tyle labilne konformacyjnie, że nie można im przypisać jednej ściśle zdefiniowanej struktury trzeciorzędowej (ryc. 7). Białka takie nazywamy białkami wewnętrznie nieustrukturalizowanymi, nieporządkowanymi lub pozbawionymi struktury trzeciorzędowej (ang. *intrinsically unstructured/disordered protein*, IUP). Analogicznie, jeśli zjawisko dotyczy tylko określonego regionu mówimy o regionach nieustrukturalizowanych (ang. *intrinsically unstructured region, IUR*). Białka tego typu są niejako zaprzeczeniem jednego z podstawowych paradygmatów biologii molekularnej, mówiącego, że za funkcję białka odpowiada ściśle zdefiniowana struktura przestrzenna. W tym przypadku, pomimo, a nawet dzięki braku ściśle zdefiniowanej struktury, poprzez dużą labilność konformacyjną białka IUP mogą pełnić bardziej złożone funkcje oddziałując w komórce z wieloma białkami. Białka IUP lub posiadające regiony IUR związane są z procesami takimi jak: proliferacja, apoptoza, powstawanie nowotworów, przekazywanie sygnału oraz regulacja transkrypcji i translacji. Właściwie można powiedzieć, że białka tego typu występują wszędzie tam, gdzie istnieje potrzeba tworzenia wielu oddziaływań między różnymi makrocząsteczkami (Tompa, 2010). Na podstawie przewidywań komputerowych przyjmuje się, że około 30% białek eukariotycznych posiada regiony IUR (Ward et al., 2004). Ponadto potwierdzono, że niektóre z takich regionów, pomimo swojej odmienności od regionów przyjmujących ściśle określona strukturę także bywają silnie konserwowane ewolucyjnie (Schlessinger et al., 2011).

Należy podkreślić, że nie ma jednej ścisłej definicji nieustrukturalizowania. Jednym z najpowszechniej stosowanych wyznaczników regionów wewnętrznie nieuporządkowanych jest brak danych doświadczalnych o strukturze trzeciorzędowej dla określonego regionu białka w bazie PDB (tzw. brakujące reszty aminokwasowe oznaczone w pliku PDB przez znacznik REMARK 465). Dotyczy to zarówno danych z dyfrakcji rentgenowskiej, gdzie reszty te nie są widoczne w ogóle, jak i techniki NMR, gdzie reszty takie cechują się dużą zmiennością współrzędnych pomiędzy alternatywnymi modelami. Ponadto istnieje szereg innych technik doświadczalnych, które pozwalają na identyfikację białek IUP. Zazwyczaj procedura wykrywania białek IUP opiera się na tym, że białka te cechują się nietypowym zachowaniem w

porównaniu do białek globularnych. Przykładowo białka takie znacznie łatwiej ulegają proteolizie. Ponadto białka IUP dają nietypowy wzór w czasie elektroforezy SDS-PAGE, bo migrują one w żelu wolniej niż by to wynikało z ich masy cząsteczkowej (Tompa, 2010).

Poza podziałem typów nieuporządkowania według użytych do ich wykrycia technik doświadczalnych stosuje się także podział funkcjonalny białek IUP. W chwili obecnej samo określenie, że białko zawiera regiony IUR, zaczyna być niewystarczające. Coraz częściej mówi się o różnych typach lub klasach białek IUP (Vucetic et al., 2003). W najprostszym przypadku białko IUP ma formę kłębka statystycznego, który ma za zadanie pełnić rolę sączka molekularnego, niczym pory gąbki, fizycznie ograniczając transport między określonymi kompartmentami komórki. Z sytuacją taką mamy do czynienia w przypadku nukleoporyn jądrowego kompleksu porowego (Denning et al., 2003). Inną podgrupą białek IUP są białka, które względnie łatwo zmieniają stopień uporządkowania w wyniku kontaktu z innymi białkami. Często obserwuje się ciekawe zjawisko polegające na tym, że przejście ze stanu nieuporządkowanego do określonej struktury wyższego rzędu uwarunkowane jest rodzajem partnera. Przykładowo ten sam region IUR białka p53 może przybierać inny kształt w zależności od białka, z którym w danej chwili oddziałuje (Uversky et al., 2008). Cechą charakterystyczną regionów IUR tego typu jest tendencja do tworzenia różnych struktur drugorzędowych, którą można wykryć za pomocą programów do przewidywania struktury drugorzędowej. Regiony te z jednej strony przewidywane są jako nieuporządkowane, a z drugiej strony przewidywanie struktury drugorzędowej lokalizuje w tym samym miejscu potencjalne helisy α. Przeciwieństwem tego typu regionów IUR są sekwencje o niskiej złożoności. W wielu białkach można zaobserwować występowanie powtarzalnych motywów złożonych jedynie z kilku typów reszt aminokwasowych. Dobrym przykładem jest C-końcowa domena polimerazy RNA II, która w białku ludzkim zawiera 52 heptapeptydowe powtórzenia o sekwencji konsensusowej YSPTSPS. Region ten jest jednocześnie nieuporządkowany i konserwowany ewolucyjnie oraz dodatkowo podlega intensywnej modyfikacji posttranslacyjnej (Egloff i Murphy, 2008). Ukoronowaniem klasyfikacji białek IUP ma być ich ontologia, nad którą trwają aktualnie intensywne prace (komunikacją ustna z Keithem Dunkerem).

W wyniku nagromadzenia danych doświadczalnych o białkach IUP powstała potrzeba stworzenia dedykowanej im bazy danych. Obecnie dostępne bazy danych to:

DISPROT – http://www.disprot.org/ –najstarsza i zarazem największa baza danych dotycząca białek IUP, w wersji 6.01 z dnia 15.10.2012 zawierała opis 684 białek IUP z 1513 regionami IUR (Vucetic et al., 2005).

MobiDB – http://mobidb.bio.unipd.it/ – baza ta zawiera dodatkowo, oprócz informacji dostępnych w DISPROT, informacje na temat brakujących reszt aminokwasowych (REMARK 465) z bazy PDB oraz przewidywania nieuporządkowania według trzech programów: ESpritz (Walsh et al., 2012), IUPred (Dosztanyi et al., 2005) i DisEMB (Linding et al., 2003a). Baza ta powstała w 2012 roku (Di Domenico et al., 2012).

IDEAL – ang. *Intrinsically Disordered proteins with Extensive Annotations and Literature*, http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/ – baza skupia się na danych z literatury ze szczególnym uwzględnieniem miejsc posttranslacyjnej modyfikacji i oddziaływania z innymi makrocząsteczkami. Duży wysiłek włożono w identyfikację i opisanie szczególnego typu regionów IUR, który cechuje się przejściem ze stanu nieuporządkowania w stan zdefiniowanej struktury trzeciorzędowej indukowanym przez oddziaływanie z inną makrocząsteczką (Fukuchi et al., 2012).

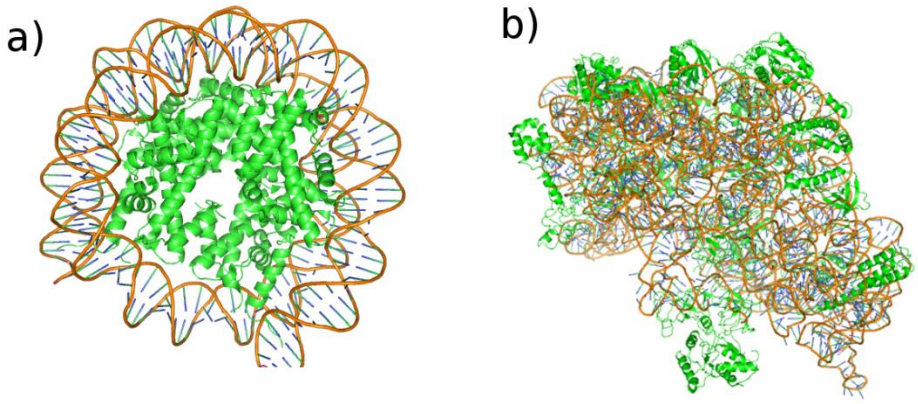
Przewidywanie regionów nieustrukturalizowanych w białkach jest jednym z głównych zagadnień omówionych w niniejszej rozprawie doktorskiej.

Ryc. 7. Przykładowe białko IUP (pdb: 2kzw) zawierające silnie nieuporządkowany koniec N. Na rycinie nałożono 10 modeli otrzymanych techniką NMR. Centralna część białka tworzy ściśle określona strukturę, natomiast koniec N (po prawej) cechuje duża swoboda. Kolorem czerwonym zaznaczono elementy struktury drugorzędowej.

3.1.5.5 Oddziaływanie białek z DNA i białek z RNA

Białka stanowią główny składnik organiczny budujący komórkę żywą. Jednak w komórce znajdują się także inne ważne makrocząsteczki. Jednym z obficiej występujących związków wielkocząsteczkowych są kwasy nukleinowe, RNA i DNA. Duża część białek ma możliwość oddziaływania z nimi. Oddziaływania te mogą mieć znacznie strukturalne, jak w przypadku białek histonowych, które tworzą „rdzeń", wokół którego nawinięty jest DNA lub funkcjonalne, gdy komponent białkowy odpowiada za określoną aktywność kompleksu na przykład katalizuje syntezę kwasów nukleinowych. Nierzadko białka i kwasy nukleinowe tworzą kompleksy, których działanie zależy na równi od składnika białkowego jak i od RNA. Przykładem może tu być rybosom, który jest olbrzymim kompleksem składającym się z blisko 100 białek i cząsteczek RNA (ryc. 8).

Oddziaływania białek z DNA lub z RNA są niezwykle istotne dla kluczowych procesów życiowych oraz umożliwiają dostęp do informacji genetycznej i jej wykorzystanie. Analiza kompleksów białko z DNA lub RNA wykazała, że za oddziaływania odpowiadają zwykle dodatnio naładowane, hydrofilowe, dostępne dla rozpuszczalnika reszty Lys i Arg oraz reszty polarne takie jak Ser, Tyr, Thr i Asn.. Ponadto regiony odpowiedzialne za odziaływanie są zazwyczaj ubogie w reszty niosące ładunek ujemny (Asp i Glu). Dodatkowo na zdolność wiązania DNA/RNA największy wpływ ma możliwość tworzenia wiązań wodorowych, wielkość reszty aminokwasowej, struktura drugorzędowa oraz hydrofobowość (Jones et al., 2001; Jones et al., 1999; Luscombe i Thornton, 2002).



Ryc. 8. Kompleksy białek z DNA i białek z RNA: a) nukleosom zbudowany z oktameru histonowego i nawiniętego na niego DNA (pdb: 2nqb); b) podjednostka 70S rybosomu (pdb: 2wrn). Kolorem zielonym zaznaczono składnik białkowy. Rysunki a i b są w innej skali.

3.2. Przewidywanie właściwości biochemicznych białek w oparciu na sekwencjach

Zazwyczaj celem badań białek jest poznanie ich funkcji i mechanizmu działania. Aby to osiągnąć, niezwykle pomocna jest znajomość struktury trzeciorzędowej białka. Niestety rozwiązanie struktury białek jest procesem żmudnym i obarczonym wysokim prawdopodobieństwem porażki. W chwili obecnej standardowo stosuje się w tym celu dwie techniki doświadczalne: krystalografię rentgenowską (88% struktur w PDB) oraz spektroskopię NMR (11,2% struktur w PDB). Inne techniki doświadczalne takie jak mikroskopia elektronowa czy małokątowe rozpraszanie promieni X lub neutronów (ang. *small angle X-rays, neutrons scattering*, SAXS lub SANS) także są stosowane, jednak liczba struktur o dobrej rozdzielczości (< 4 Å) otrzymanych tymi metodami jest bardzo mała (poniżej 1% bazy PDB). Wśród nowych technik warto wspomnieć o dyfrakcji rentgenowskiej z wykorzystaniem laserów opartych na swobodnych elektronach (ang. *X-ray Free Electron Laser*, X-FEL), które emitują impulsową wiązkę światła laserowego o bardzo dużej mocy. Zaletą tego rozwiązania jest to, że można wykorzystać kryształy znacznie mniejszych rozmiarów niż ma to miejsce w standardowej krystalografii rentgenowskiej. W chwili obecnej pojawiają się pierwsze struktury otrzymane tą techniką np. struktura proteazy cysteinowej, katepsyny B z *Trypanosoma brucei* (Redecke et al., 2012). Niemniej jednak istnieje olbrzymi rozdźwięk między ilością danych sekwencyjnych i strukturalnych (przykładowo baza nr zawiera ponad 20 milionów sekwencji białkowych, natomiast w bazie PDB znajduje się około 90 tysięcy struktur). Mając na uwadze doświadczenia Anfinsena wydaje się, że powinno być możliwe przewidywanie struktury trzeciorzędowej opierając się jedynie na sekwencji. Niestety do chwili obecnej nie udało się wystarczająco dokładnie określić reguł, które decydują o zwijaniu się białek. Precyzyjne przewidywanie struktury białek z dokładnością atomową za pomocą metod obliczeniowych ciągle pozostaje nieosiągalne. Można jedynie mówić o sporadycznych sukcesach np. przy wykorzystaniu programu Rosetta (Rohl et al., 2004), ale i w tym przypadku mamy do czynienia raczej z małymi, jednodomenowymi białkami, których wielkość nie przekracza 100 reszt aminokwasowych. W tym miejscu zaznaczyć trzeba, że prócz przewidywania struktury trzeciorzędowej, także inne właściwości białek takie jak struktura drugorzędowa, obecność helis transbłonowych, sekwencji sygnalnych, mostków dwusiarczkowych i wiele innych same w sobie niosą dużo informacji, która często pozwala określić funkcję czy mechanizm działania białka. Z

tego względu programy przewidujące takie właściwości są niezwykle użyteczne. W najprostszej formie programy takie działają w oparciu na samej sekwencji aminokwasowej. W tym przypadku dla każdego aminokwasu lub grupy aminokwasów sąsiadujących z rozpatrywaną pozycją (tzw. okna uwzględniającego od n-m do n+m aminokwasów od pozycji n w sekwencji, gdzie m to szerokość okna) na podstawie parametrów biochemicznych przypisanych do poszczególnych reszt aminokwasowych otrzymanych na drodze doświadczalnej liczona jest określona metryka, która porównywana jest z danymi znanymi do tej pory. Przykładowo, najprostsze metody do przewidywania struktury drugorzędowej opierały się na częstości występowania reszt hydrofobowych, hydrofilowych, kwaśnych, zasadowych, wielkości grupy funkcyjnej czy obecności grup aromatycznych w regionach o określonej strukturze drugorządowej. Dane o właściwościach fizykochemicznych zebrano w bazie danych o nazwie AAindex (Kawashima et al., 2008). W chwili obecnej baza ta zawiera ponad 500 różnego rodzaju metryk określających względne występowanie określonej właściwości biochemicznej w zależności od typu reszty aminokwasowej. Należy podkreślić, że baza AAIndex zawiera pewną nadmiarowość, tzn. dozwolone jest umieszczenie kolejnej skali np. skali hydrofobowości o ile została ona wyznaczona doświadczalnie (różnice między wynikami mogą wynikać m.in. z innego środowiska, odmiennego urządzenia pomiarowego, innego podejścia doświadczalnego, lub innej grupy białek, która została wykorzystana w czasie badań). Istnieje bardzo duża grupa programów, których działanie opiera się na wykorzystaniu optymalnej kombinacji metryk pochodzących z bazy AAindex w celu przewidywania określonej cechy białka (np. (Cai i Lu, 2008), (Lu et al., 2009), (Huang et al., 2011) i wiele innych). Standardowy schemat budowy takiego programu wygląda następująco. Najpierw z bazy AAindex wybiera się te metryki, które mają największe znaczenie dla rozpatrywanego problemu. W tym celu można użyć technik pozwalających na statystyczne określenie wpływu określonej cechy na dany problem (ang. *feature selection*). Następnie za pomocą uczenia maszynowego (wykorzystując np. algorytm ANN lub SVM) konstruuje się model komputerowy, który najlepiej tłumaczy obserwowaną cechę w zbiorze testowym.

Bardziej złożone programy do przewidywania cech białek wykorzystują dodatkowo informację ewolucyjną. Zasadniczo idea tego etapu sprowadza się do identyfikacji sekwencji białek homologicznych np. za pomocą programu PSI-BLAST (czasem mogą być do tego użyte inne programy). Na podstawie przyrównania MSA tworzony jest profil PSSM, który można

wykorzystać jako dodatkowe źródło informacji o sekwencjach homologicznych. Poprawa jakości przewidywania wiąże się z tym, że dzięki użyciu wielu sekwencji wprowadzamy większą ilość informacji, która może być uwzględniona w modelu komputerowym. Ponadto budując przyrównania MSA wskazujemy, które aminokwasy są ważniejsze od innych (wnioskowanie jest oparte na podstawie konserwacji). Schemat działania tego rodzaju programów wygląda następująco. W pierwszym etapie uruchamiany jest program do przeszukiwania bazy znanych sekwencji (najczęściej jest to program PSI-BLAST i dowolna duża baza danych tj. nr lub UniProt). Następnie na podstawie zidentyfikowanych sekwencji tworzone jest przyrównanie MSA i profil PSSM. Czasem dochodzi jeszcze kolejny etap polegający na przekształceniu wyników programu PSI-BLAST na modele HMM, które mogą być wykorzystane do przeszukania baz danych modeli HMM (np. przy użyciu programu HHsearch). Posiadając profil PSSM lub model HMM można go użyć jako dane wejściowe do algorytmu uczenia maszynowego (np. techniki SVM).

Oczywiście nic nie stoi na przeszkodzie, aby oba typy programów połączyć i utworzyć program, który integruje informację o właściwościach fizykochemicznych reszt aminokwasowych z informacją ewolucyjną. Należy jednak zaznaczyć, że największą poprawę wyników uzyskuje się na etapie wprowadzania informacji ewolucyjnej, dlatego też przeszukiwanie baz sekwencyjnych za pomocą programu PSI-BLAST stało się standardową techniką i większość programów przewidujących właściwości białek na podstawie sekwencji zawiera ten etap.

3.3. Meta-metody do przewidywania właściwości białek

W miarę jak ilość różnych metod do przewidywania określonych właściwości białek wzrastała, zaczęły pojawiać się zarówno korzyści jak i pewne problemy wynikające z tego stanu rzeczy. Z jednej strony problematyczne okazało się wiarygodne porównanie nowych metod z istniejącymi (np. ze względu na specyfikę zbiorów testowych oraz fakt, że zbiory testowe nowych metod mogą częściowo nachodzić na zbiory danych wykorzystane w trakcie uczenia starszych metod, itp.). Z drugiej strony, powstanie kilku metod próbujących rozwiązać ten sam problem, ale w odmienny sposób (np. wykorzystane są inne techniki uczenia maszynowego, inne zbiory testowe), daje szansę na skonstruowanie nowej metody, która połączy ich wyniki i da

wynik lepszy niż jej części składowe. Metody takie nazywa się meta-metodami. W najprostszym przypadku mamy do czynienia z sytuacją, w której nowa metoda wykorzystuje jednie metody przewidujące określoną właściwość np. na podstawie wyniku różnych metod służących do przewidywania struktury drugorzędowej konstruujemy „nadrzędną" meta-metodę do przewidywania struktury drugorzędowej. Oczywiście zwykle sytuacja jest bardziej skomplikowana i do przewidywania wybranej właściwości można wykorzystać przewidywania innych właściwości, o ile są one skorelowane z tą, którą chcemy przewidywać. Dla przewidywania struktury drugorzędowej znaczenie ma przewidywanie m.in. dostępności reszt aminokwasowych dla rozpuszczalnika oraz regionów wewnętrznie nieuporządkowanych. Właściwie na tym etapie należałoby wspomnieć, że trudno jest wyznaczyć jednoznaczną granicę, powyżej której programy należy uznać za meta-metody, ponieważ np. przewidywanie struktury drugorzędowej (najczęściej za pomocą programu PSIPRED) jest elementem składowym praktycznie każdego programu przewidującego inne właściwości białek, a sam program PSIPRED używa programu PSI-BLAST.

Kolejną otwartą kwestia pozostaje sposób, w jaki dokonuje się integracji metod składowych meta-metody w finalny model. W najprostszym przypadku można zbudować konsensus oparty na zwykłej średniej arytmetycznej czy zasadzie większości. Często jednak stosując tak proste rozwiązanie otrzymuje się model, który nie poprawia ogólnej jakości końcowego wyniku. W efekcie otrzymujemy metodę, która działa średnio dobrze, ale nie lepiej niż najlepsza składowa. W takim przypadku należy uciec się do bardziej skomplikowanych modeli jak średnia ważona czy uczenie maszynowe, które maja na celu optymalne połączenie metod składowych tak, aby meta-metoda wzmacniała prawdziwy sygnał przy jednoczesnym tłumieniu fałszywych sygnałów. Stosując tego typu techniki względnie łatwo osiągnąć kilkuprocentową poprawę przewidywania dowolnej właściwości białka. Przykładami takich meta-metod może być metoda MetaMQAP służąca do przewidywania dokładności modeli struktury przestrzennej białek (Pawlowski et al., 2008) czy prezentowana w tej rozprawie metoda do przewidywania rejonów wewnętrznego nieuporządkowania w białkach (Kozlowski i Bujnicki, 2012).

3.4. Białka odpowiedzialne za modyfikację końca 3′ mRNA

Podczas tworzenia wszelkich programów bioinformatycznych należy pamiętać, że ich użyteczność w praktyce zależy przede wszystkim od tego, na ile pomóc mogą badaczom prowadzącym analizy doświadczalne – w interpretacji już posiadanych wyników lub do generowania nowych hipotez roboczych, które można później zweryfikować doświadczalnie. Dlatego zawsze poza sprawdzeniem statystycznej poprawności przewidywań powinno się również zweryfikować użyteczność tworzonych narzędzi. W wyniku takiej weryfikacji programy często zostają rozbudowane pod kątem dodatkowych funkcjonalności, które pomagają zinterpretować wynik programu (np. poprzez dodanie określonych elementów interfejsu, albo zmianę szaty graficznej w celu poprawienia czytelności), a które ze względu na sam algorytm programu mogą wydawać się nieistotne, ale mogą mieć kluczowe znaczenie dla losów algorytmu: czy jego konkretna implementacja będzie używana w praktyce, czy nie.

Aby wykazać praktyczną użyteczność wszystkich prezentowanych w niniejszej rozprawie programów komputerowych, zostały one wykorzystane do przeprowadzenia analizy sekwencji ludzkich białek odpowiedzialnych za modyfikację końca 3′ mRNA. Białka te tworzą kompleks, którego rdzeń zbudowany jest z około 30 białek odpowiedzialnych za rozpoznawanie, cięcie i poliadenylację końca 3′ mRNA (ryc. 9) (Mandel et al., 2008; Millevoi i Vagner, 2010; Yang i Doublie, 2011). Kompleks ten jest niezwykle dynamiczny w czasie i przestrzeni (poszczególne białka przyłączają się na określonych etapach i po spełnieniu swojej funkcji zastępowane są przez inne), dlatego ustalenie liczby tworzących go cząsteczek jest arbitralne. Badania oparte na analizie danych pochodzących ze spektrometrii mas pozwoliły zidentyfikować około 85 białek, które można znaleźć w oczyszczonym ekstrakcie zawierającym wspomniany kompleks (Shi et al., 2009). Wśród dodatkowych, nienależących do rdzenia kompleksu 50 białek, znajdują się m.in. czynniki transkrypcyjne, białka biorące udział w procesie składania transkryptów (ang. *splicing*) czy białka zaangażowane w naprawę DNA. Najprawdopodobniej łączą one działanie kompleksu z innymi procesami biologicznym. W obrębie rdzenia kompleksu głównego można wyróżnić następujące podkompleksy:

- kompleks czynników stymulacji cięcia (ang. *cleavage stimulation factor,* CstF) złożony z białek CstF50 (CSTF1), CstF77 (CSTF3), CstF64 (CSTF2) lub jego formy Tau CstF64 (CSTF2T),

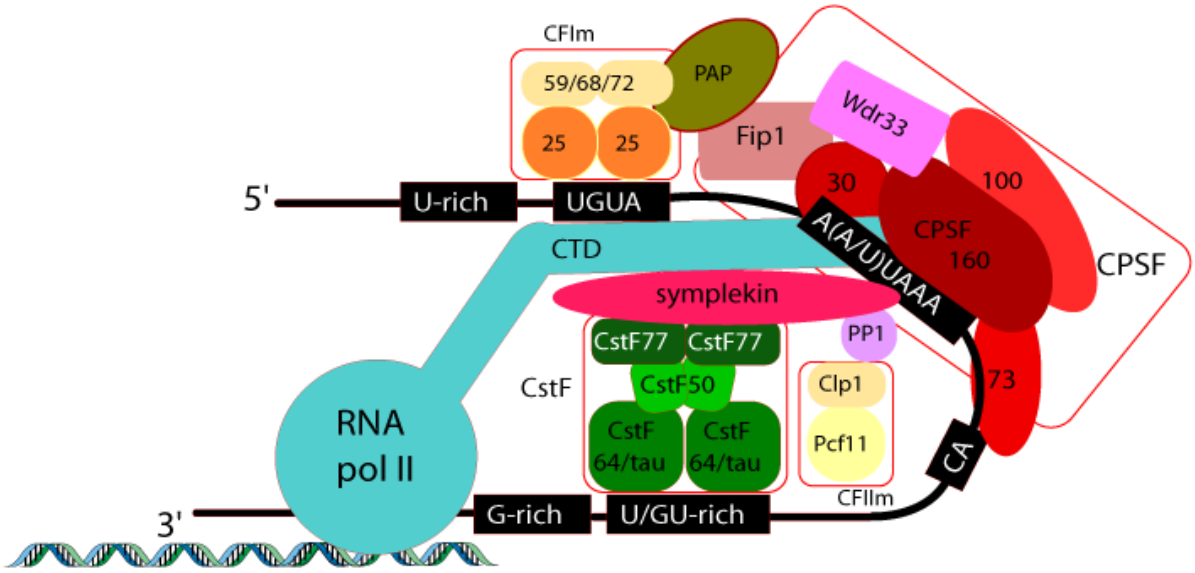
- kompleks czynników specyficzności cięcia i poliadenylacji (ang. *cleavage and polyadenylation specificity factor*, CPSF) złożony z białek CPSF160, CPSF100, CPSF73, CPSF30, Fip1 oraz Wdr33,
- ssaczy kompleks czynników cięcia typu I (ang. *mammalian cleavage factor I,* CF Im) złożony z białek CF Im 68 (CPSF6), CF Im 59 (CPSF7), CF Im 25 (CPSF5),
- ssaczy kompleks czynników cięcia typu II (ang. *mammalian cleavage factor II,* CF IIm) złożony z białek Clp1, Pcf11.

Ogólna zasada działania kompleksu modyfikującego koniec 3′ mRNA jest znana i polega na rozpoznaniu konserwowanej ewolucyjnie sekwencji sygnalnej A(A/U)UAAA zlokalizowanej powyżej miejsca cięcia oraz kasety bogatej w uracyl i guaninę (ang. *U/GU-rich element*) położonej 10-30 nukleotydów poniżej miejsca cięcia. Następnie po przyłączeniu poszczególnych składników kompleksów CPSF i CstF oraz innych białek pomocniczych dochodzi do specyficznego cięcia, które katalizuje domena endonukleolityczna białka CPSF73. W ostatnim etapie dochodzi do poliadenylacji, czyli dołączenia do końca 3′ „ogona" zbudowanego z wielu reszt adenozyny (za proces ten odpowiada polimeraza poli(A)). Długość ogona ma decydujące znaczenie dla transportu mRNA z jądra komórkowego, jego stabilności i translacji. Wraz z upływem czasu ogon poli(A) skraca się i kiedy jest już odpowiednio krótki cały mRNA ulega enzymatycznej degradacji (Guhaniyogi i Brewer, 2001).

Kompleks odpowiedzialny za modyfikacje końca 3′ mRNA jest konserwowany ewolucyjnie od drożdży po człowieka (Darmon i Lutz, 2012). Początkowo może dziwić fakt, że za ten względnie prosty proces odpowiedzialne jest tak wiele białek. Przyczyną tego stanu rzeczy jest prawdopodobnie to, że proces ten musi być ściśle kontrolowany i ponieważ wszelkie nieprawidłowości w jego przebiegu skutkują licznymi zaburzeniami i chorobami (Danckwardt et al., 2008).

Należy podkreślić, że choć znamy ogólny schemat przebiegu modyfikacji końca 3′ mRNA oraz podstawowe składniki odpowiedzialnego za ten proces kompleksu, ciągle nie jesteśmy w stanie scharakteryzować go w pełni na poziomie molekularnym. Struktury dużej części białek kompleksu lub ich domen rozwiązano metodą krystalografii rentgenowskiej. Ciągle jednak istnieje również wiele białek, co do których posiadamy jedynie szczątkowe informacje. Za pomocą stworzonych w ramach niniejszej rozprawy doktorskiej programów dokonano analizy domen oraz regionów wewnętrznie nieuporządkowanych białek kompleksu

odpowiedzialnego za modyfikacje końca 3′ mRNA. Ponadto zbudowano modele homologiczne wszystkich białek kompleksu. Uzyskane wyniki istotnie przyczyniają się do poszerzenia wiedzy na temat struktury tego kompleksu i pomogą zarówno w interpretacji istniejących danych doświadczalnych jak i w planowaniu nowych analiz.



Ryc. 9. Zgrubny schemat rdzenia kompleksu białek odpowiedzialnych za modyfikację końca 3′ mRNA.

4. Cel rozprawy

Głównym celem niniejszej rozprawy doktorskiej było stworzenie zintegrowanego serwisu bioinformatycznego do analizy białek, ze szczególnym uwzględnieniem metod do przewidywania regionów wewnętrznie nieuporządkowanych oraz metod do przewidywania domen białkowych na podstawie sekwencji aminokwasowej. Jako punkt startowy rozprawy doktorskiej posłużyła wstępna wersja metaserwera do rozpoznawania zwoju opracowana ponad 10 lat temu w laboratorium prof. Bujnickiego (Kurowski i Bujnicki, 2003). Program ten został znacznie zmodyfikowany i rozbudowany przez dodanie nowych metod. W chwili obecnej serwis internetowy umożliwia jednoczesne uruchomienie ponad 100 programów bioinformatycznych, które pozwalają przeprowadzić wszechstronną analizę sekwencji badanego białka pod kątem takich cech jak struktura drugorzędowa, występowanie helis transbłonowych, sekwencji sygnalnych, splecionych helis, mostków dwusiarczkowych, domen, regionów wewnętrznie nieuporządkowanych oraz dostępności reszt aminokwasowych dla rozpuszczalnika. Ponadto możliwe jest porównanie wyników metod przewidujących odziaływanie białek z DNA i RNA oraz szczegółowa analiza najbliższych homologów o znanej strukturze wykrytych przez programy do rozpoznawania zwoju. Na podstawie struktury wykrytych homologów automatycznie budowane są modele homologiczne, następuje klasyfikacja według podobieństwa do struktur z bazy danych SCOP oraz określane jest czy dana sekwencja należy do białka enzymatycznego. Wyniki pochodzące z programów określonej kategorii są przedstawione w jednorodnym formacie w celu ułatwienia interpretacji.

W czasie realizacji prezentowanych badań szczególny nacisk położono na przewidywanie regionów wewnętrznie nieuporządkowanych oraz na przewidywanie domen białkowych. W efekcie opracowano nowe algorytmy pozwalające na bardziej wiarygodne przewidywanie tych cech białka. Pierwszym celem badawczym było stworzenie nowego, lepszego programu do przewidywania regionów wewnętrznie nieuporządkowanych. Metody komputerowe tego typu są cennym źródłem biologicznie istotnych informacji, ponieważ wykazano, że brak uporządkowania jest często występującą cechą białek związanych z proliferacją, apoptozą czy powstawaniem nowotworów (Haynes et al., 2006b). Ponadto przewidywanie braku uporządkowania struktury białka jest ważne ze względów technicznych. Bardzo często rozwiązanie struktury białka metodą krystalografii rentgenowskiej nie jest możliwe, ponieważ nie można otrzymać kryształów wystarczającej jakości. Jedną z głównych

przyczyn tego zjawiska jest obecność regionów wewnętrznie nieuporządkowanych w docelowym białku. W związku z tym ustalenie ich obecności (np. na podstawie przewidywania komputerowego) i usunięcie ich przed krystalizacją często umożliwia rozwiązanie struktury pozostałej części białka (Dosztanyi et al., 2007). W chwili obecnej komputerowe przewidywanie braku uporządkowania jest standardową procedurą we wszystkich większych centrach genomiki strukturalnej. W momencie rozpoczęcia tej części pracy istniało już ponad 20 programów tego typu, dlatego naturalnym krokiem było stworzenie meta-metody, której działanie oparte zostało na wynikach pochodzących z istniejących już programów. Sama integracja wyników była zadaniem na tyle skomplikowanym, że w kolejnych wersjach programu wykorzystano oprócz prostych metod statystycznych także metody uczenia maszynowego (sztuczne sieci neuronowe i algorytmy genetyczne). Wynikiem tej części badań jest cała grupa programów przewidujących rejony sekwencji białka pozbawione struktury trzeciorzędowej, które różnią się od siebie sposobem tworzenia konsensusu (od prostych metod takich jak średnia ważona po bardziej złożone algorytmy genetyczne) oraz pod względem cech na podstawie, których tworzone są przewidywania (początkowa wersja uwzględniała jedynie wyniki innych programów do przewidywania braku uporządkowania, natomiast wersja końcowa uwzględniała dodatkowo przewidywanie struktury drugorzędowej, dostępności reszt aminokwasowych dla rozpuszczalnika oraz pokrycie w przyrównaniu do najbliższego homologa wykrytego przez programy do rozpoznawania zwoju). Należy podkreślić, że programy do przewidywania braku struktury trzeciorzędowej, które powstały w ramach niniejszej rozprawy doktorskiej, zostały bardzo intensywnie przetestowane nie tylko przez autora, ale także w trakcie międzynarodowych „zawodów" CASP (ang. ***C**ritical **A**ssessment of Techniques for Protein **S**tructure **P**rediction*). W przeciągu dwóch kolejnych edycji konkursu CASP8 i CASP9 metody te zdobywały najlepsze miejsca w swojej kategorii. Praca szczegółowo opisująca tą część badań ukazała się w *BMC Bioinformatics* w maju 2012 roku (Kozlowski i Bujnicki, 2012).

Drugim celem badawczym niniejszej rozprawy doktorskiej było stworzenie programu do przewidywania domen białkowych na podstawie sekwencji aminokwasowej. Program ten w pierwszej kolejności klasyfikuje białko jako jednodomenowe lub wielodomenowe. W przypadku białek wielodomenowych proponuje lokalizację granic domen. Program przewiduje występowanie domen opierając się na takich cechach jak entropia, hydrofobowość, przewidywana struktura drugorzędowa, dostępność reszt aminokwasowych dla rozpuszczalnika,

występowanie regionów wewnętrznie nieuporządkowanych oraz obecność bliskich homologów w bazach domen białkowych CATH i PFAM. Od strony algorytmicznej program ten opiera swoje działanie na zastosowaniu maszyny wektorów nośnych (ang. *support vector machine, SVM*).

W ostatnim etapie badań stworzone programy wykorzystano do analizy białek odpowiedzialnych za modyfikację końca 3′ mRNA. Jest to grupa około 30 białek tworzących kompleks, którego funkcją jest cięcie i poliadenylacja końca 3′ mRNA. W ramach tej analizy zbadano występowanie domen i regionów wewnętrznie nieuporządkowanych oraz zbudowano modele homologiczne białek kompleksu.

5. Materiały i metody

W ramach rozprawy doktorskiej opisane zostały trzy programy:

a) metaserwer GeneSilico – **https://genesilico.pl/meta2/** – serwis do rozpoznawania zwoju oraz analizy innych właściwości białek
b) GeneSilico MetaDisorder – **http://genesilico.pl/metadisorder/** – serwis do przewidywania wewnętrznie nieuporządkowanych regionów białka
c) DomainSVM – program do przewidywania domen białkowych

Dwa ostatnie programy zostały zintegrowane z metaserwerem GeneSilico. Wszystkie programy i skrypty zostały napisane w języku programowania Python.

Dodatkowo wyniki analizy białek odpowiedzialnych za modyfikację końca 3′ mRNA zostały udostępnione w formie interaktywnej bazy danych dostępnej pod adresem **http://genesilico.pl/mrna3db**.

5.1. Sprzęt komputerowy

Ze względu na złożoność obliczeniową, każdy z serwisów używa innych zasobów komputerowych:

a) metaserwer GeneSilico

   Baza danych, część programów 32 bitowych oraz interfejs internetowy zainstalowane są na komputerze o następującej specyfikacji:

   4 CPU x 2,6 GHz (Dual Core AMD Opteron Processor 285); pamięć: 16 GB RAM; przestrzeń dyskowa: 160GB; system: Debian 4.0.

   Ze względu na wydajność większość obliczeń wykonywana jest na innej maszynie:

   48 x 2,2 GHz (AMD Opteron Processor 6174); pamięć: 96 GB RAM; przestrzeń dyskowa: 8,7 TB; system: Debian 6.0.

b) GeneSilico MetaDisorder

   Baza danych, część programów niewymagających dużej mocy obliczeniowej i interfejs internetowy zainstalowane są na wirtualnej maszynie XEN:

   2 x 2,2 GHz (AMD Opteron(tm) Processor 6174); pamięć: 2 GB RAM; przestrzeń dyskowa: 20 GB; system: Debian 4.0.

   Pozostałe programy uruchamiane są na 48 procesorowej maszynie, której specyfikacje podano w poprzednim punkcie

c) DomainSVM

Komputer klasy PC: *16 x 2,4 GHz (Intel Xeon CPU E5620); pamięć: 24 GB RAM; przestrzeń dyskowa 1,5 TB; system: Ubuntu 12.04*.

5.2. Bazy danych

Wszystkie programy wykorzystują w celu gromadzenia i zarządzania danymi MySQL – system zarządzania relacyjnymi bazami danych oparty na otwartej licencji GPL. Dodatkowo wiele programów wymaga do działania biologicznych baz danych. W chwili obecnej wykorzystywane bazy to:

- **nr** (NCBI) – znane i przewidywane nieopatentowane sekwencje białkowe pochodzące ze zbiorów GenBank (translacje sekwencji kodujących), PDB, SwissProt, PIR i PRF. W grudniu 2012 roku baza zawierała 21,5 miliona sekwencji (12 BG danych). Baza danych aktualizowana jest w każdy wtorek.
- **nr90** (NCBI) – baza zawierająca sekwencje reprezentatywne dla sekwencji białek z bazy nr na poziomie identyczności co najmniej 90%. Klastrowanie przeprowadzane jest programem CD-HIT (Li i Godzik, 2006). W grudniu 2012 roku baza zawierała 12 milionów sekwencji (6,3 GB danych). Baza jest generowana automatycznie zaraz po aktualizacji bazy nr.
- **PDB** – struktury białkowe rozwiązane metodami doświadczalnymi, w grudniu 2012 baza zawierała ponad 86 tysięcy struktur (67 GB danych). Baza jest aktualizowana codziennie o północy.
- **CULLPDB** – baza sekwencji pochodzących ze struktur PDB pobrana z serwisu PISCES (Wang i Dunbrack, 2003), Ze struktur o rozdzielczości mniejszej niż 3 Å i współczynniku *R-factor* poniżej 1.0 wybierane są sekwencje na poziomie identyczności co najmniej 90%.. Baza aktualizowana jest w każdy poniedziałek.
- **PDB70** – baza modeli HMM zawierająca modele reprezentatywne dla białek z bazy PDB na poziomie identyczności sekwencji co najmniej 70%, w grudniu 2012 roku baza zawierała 27 tys. struktur. Baza aktualizowana jest w każdy wtorek.
- **DSSP** – baza zawierająca strukturę drugorzędową i RSA dla wszystkich struktur z bazy PDB wygenerowana za pomocą programu DSSP (Kabsch i Sander, 1983). Aktualizacja

następuje codziennie o północy.

- **Pfam** – ang. ***p**rotein **fam**ily database* – baza rodzin białkowych reprezentowana w formie modeli HMM, w grudniu 2012 zawierała 11 milionów sekwencji sklasyfikowanych do 13,5 tysiąca rodzin. Aktualność bazy danych sprawdzana jest w każdy piątek.
- **CDD** – ang. *Conserved Domain Database* – Baza Domen Konserwowanych, baza NCBI reprezentowana w formie pozycyjnie specyficznych macierzy oceniających (PSSM), w grudniu 2012 zawierała 28 tysięcy macierzy.
- **UniProtKB/Swiss-Prot** – baza danych sekwencji białkowych, w grudniu 2012 zawierała ponad 450 tysięcy sekwencji, aktualizacja bazy odbywa się w każdy wtorek.
- **Uniref90 –** pochodna bazy danych UniProtKB, zawiera sekwencje reprezentacyjne na poziomie identyczności sekwencji co najmniej 90%. Klastrowanie przeprowadzane jest programem CD-HIT (Li i Godzik, 2006). Aktualność bazy sprawdzana co miesiąc.
- **CATH** – baza domen białkowych, w wersji 3.5 z września 2011 roku 51 tysiącom sekwencji struktur pochodzących z PDB przypisano ponad 170 tysięcy domen.

Oprócz wymienionych wyżej biologicznych baz danych część programów wymaga specyficznych baz, które dostarczane i aktualizowane są przez autorów.

5.3. Oprogramowanie

5.3.1. Oprogramowanie do wizualizacji danych

W czasie testowania i interpretacji wyników poszczególnych programów niezbędne było wykorzystanie programów pozwalających na wizualizację plików w formacie PDB. Programy te zostały wykorzystane także do wykonania rycin umieszczonych w rozprawie. Programy te to:

- PyMOL (DeLano, 2002);
- UCSF Chimera (Pettersen et al., 2004);
- Swiss PDB Viewer (Guex i Peitsch, 1997).

5.3.2. Wykorzystane metody uczenia maszynowego

W ramach prezentowanych badań wykorzystano następujące metody statystyczne oraz metody uczenia maszynowego:

- sztuczne sieci neuronowe (ANN) zaimplementowane za pomocą biblioteki FANN (Nissen, 2003);
- algorytmy genetyczne zaimplementowane za pomocą biblioteki PyEvolve (Butterfield et al., 2004);
- maszyna wektorów nośnych (SVM) zaimplementowane za pomocą biblioteki LIBSVM (Chang i Lin, 2011).

Ponadto część analiz statystycznych i eksploracja danych była przeprowadzona za pomocą programu STATISTICA 7 z dodatkowym modułem DATA MINER (StatSoft, Inc. Tulsa, OK, USA).

5.3.3. Inne oprogramowanie i biblioteki pomocnicze

Interfejs internetowy metaserwera do rozpoznawania zwoju używa dostępnego bezpłatnie serwera aplikacji Zope (Z Object Publishing Environment) napisanego w języku programowania Python. Umożliwia on wydajne tworzenie systemów zarządzania treścią, portali internetowych oraz aplikacji webowych. Użyta wersja to Zope 2.10. W przypadku serwisów do przewidywania domen i braku uporządkowania wykorzystano moduł mod_python stanowiący rozszerzenie w języku Python do otwartego serwera HTTP Apache.

5.3.4. Programy i serwisy internetowe stanowiące część metaserwera do rozpoznawania zwoju

Metaserwer GeneSilico (https://genesilico.pl/meta2/) jest jednym z podstawowych narzędzi ułatwiających modelowanie homologiczne. Umożliwia on znalezienie optymalnych szablonów – białek o znanej strukturze – dla docelowej sekwencji aminokwasowej. W momencie publikacji w 2003 roku (Kurowski i Bujnicki, 2003) serwis pozwalał na automatyczne uruchomienie kilkunastu metod, w tym ośmiu metod do rozpoznawania zwoju, trzech metod do przewidywania struktury drugorzędowej, trzech metod do przewidywania helis transbłonowych oraz programu HMMER (Finn et al., 2011) do przeszukiwania bazy danych PFAM. Dla każdego

zidentyfikowanego szablonu tworzone były automatycznie modele za pomocą programu SCWRL (Dunbrack, 1999). W miarę rozwoju serwisu dodawane były kolejne programy, których liczba w roku 2013 przekroczyła 100. W chwili obecnej metaserwer GeneSilico można traktować jako jedno z najważniejszych narzędzi bioinformatycznych do analizy sekwencji białkowej. Serwis ten cieszy się dużym zainteresowaniem wśród społeczności naukowej (ponad 2300 zarejestrowanych użytkowników, ponad 32 tysiące zapytań). Istnieją tylko dwa inne serwisy o zbliżonej charakterystyce, ale każdy z nich ma swoje wady i zalety. Serwisy te to PredictProtein (Rost i Liu, 2003) oraz do niedawna metaserwer BioInfoBanku (Bujnicki et al., 2001), który został wyłączony w sierpniu 2012 z powodu braku finansowania ze źródeł publicznych.

Pod względem technicznym programy zintegrowane z metaserwerem GeneSilico możemy podzielić na dwie główne kategorie. Są to programy zainstalowane lokalnie oraz programy uruchamiane zdalnie (inne serwisy internetowe). Te ostatnie dodatkowo można podzielić na dwie podgrupy: serwisy szybkie (wynik jest prezentowany w formie strony internetowej w ciągu kilku sekund po uruchomieniu) oraz wolne (wynik jest odsyłany na podany adres email po dłuższym czasie). Od strony praktycznej najwygodniejsza jest instalacja programów lokalnie, mimo że programy te czasem wymagają dużych zasobów systemowych i mogą zależeć od innych programów lub baz danych. Zwykle, jeśli uda się już zainstalować program, działa on bezproblemowo, a ewentualne problemy można łatwo zdiagnozować i rozwiązać programistycznie. Inaczej sytuacja wygląda w przypadku serwisów internetowych, ponieważ nie mamy kontroli nad wieloma aspektami ich działania. Po pierwsze serwis może być chwilowo niedostępny, po drugie jego sposób działania może ulec zmianie (np. może nastąpić zmiana parametrów formularza strony internetowej), po trzecie część zapytań do serwisu może zakończyć się niepowodzeniem (np. metoda nie odsyła wyniku na wskazany email, adres IP jest blokowany z powodu zbyt dużej liczby zapytań) oraz wiele innych. Wszystkie te wyjątki należy uwzględnić w skryptach obsługujących zdalne metody. W wątpliwych przypadkach niezbędne jest ponowne wysłanie zapytania (po upływie określonego czasu bez odpowiedzi).

5.3.4.1. Programy do przewidywania struktury drugorzędowej

Choć pierwotnie metaserwer GeneSilico zawierał jedynie trzy programy tej kategorii (PSIPRED (McGuffin et al., 2000), SAM-T02 (Karplus et al., 2003) i PROF (Ouali i King,

2000)), w chwili obecnej możliwe jest otrzymanie wyników z ponad 20 programów. Na ich podstawie budowany jest prosty konsensus w formie średniej arytmetycznej. Wyniki wszystkich programów zredukowane są do trzyliterowego alfabetu: H (helisa α), E (wstęga β) i „–” (inne). Dostępne programy to:

- HMMSTR – program opiera swoje działanie na modelach HMM (Bystroff et al., 2000),
- SSPro4 – program opiera swoje działanie na profilach otrzymanych za pomocą programu PSI-BLAST, które są następnie przetwarzane przez sztuczną sieć neuronową (ANN) (Pollastri et al., 2002),
- JNET – algorytm ANN oparty na przyrównaniu wielu sekwencji (Cole et al., 2008),
- GOR – opiera się na statystyce Bayesa, teorii informacji i przyrównaniu sekwencji z programu PSI-BLAST (Kloczkowski et al., 2002),
- FMD – struktura drugorzędowa przewidywana jest za pomocą metody SVM, która opiera swoje działanie na podobieństwie strukturalnym do biblioteki fragmentów białek z PDB (Cheng et al., 2005a),
- CDM – dla fragmentów o dużym podobieństwie struktura drugorzędowa przewidywana jest za pomocą programu FMD, natomiast pozostałe fragmenty są oceniane przez program GOR (Cheng et al., 2007),
- PROTEUS2 – używa programów JNET, PSIPRED i SABLE, których wynik przetwarzany jest na zasadzie głosowania większości (Montgomerie et al., 2008),
- *SPARROW i SPARROW – profile z programu PSI-BLAST przetwarzane są przez metody liniowej regresji wieloczynnikowej (ang. *multiple linear regression*), różnica między programami polega na tym, że SPARROW wykorzystuje model kaskadowy (najpierw sprawdzane jest czy reszta reprezentuje kolejno „H”, „E” i „-”), w *SPARROW problem jest reprezentowany w postaci 3 oddzielnych klasyfikatorów, których wynik scalany jest w końcowym etapie (Bettella et al., 2012),
- SABLE – poprawa jakości przewidywań algorytmu ANN opiera się wykorzystaniu profili z programu PSI-BLAST oraz dodatkowo przewidywań dostępności reszt aminokwasowych dla rozpuszczalnika (Adamczak et al., 2005),
- PROF PHD – jeden z pierwszych programów oparty na algorytmie ANN i profilach z programu PSI-BLAST (Rost i Sander, 1993),
- PORTER – program opiera się na dwuwarstwowej, dwukierunkowej, rekurencyjnej sieci

neuronowej (Pollastri i McLysaght, 2005),

- NetSurfP – wielowarstwowy algorytm ANN oparty na profilach z program PSI-BLAST (Petersen et al., 2009),
- RaptorX-SS8 – korzysta z algorytmu warunkowych pól neuronowych (ang. *conditional neural fields*, CNF), siła tego algorytmu polega na tym, że pod uwagę brany jest także tzw. kontekst czyli przewidywania dla sąsiednich regionów, algorytm ten można traktować jako uogólnienie modeli HMM oparte na grafach (Wang et al., 2011),
- PSSpred – łączy w sobie 7 algorytmów ANN, które różnią się sposobem reprezentacji danych wejściowych i parametrami pochodzącymi z przyrównań wielu sekwencji z programu PSI-BLAST (Zhang, 2012),
- SPINE – wykorzystuje algorytm ANN, który bierze pod uwagę obok profili z program PSI-BLAST, także takie cechy jak hydrofobowość, objętość, polarność oraz punkt izoelektryczny (Dor i Zhou, 2007),
- SPINE X – wielowarstwowy model przewidywania struktury drugorzędowej bazujący na wyniku programu SPINE, przewidywaniu kątów torsyjnych oraz dostępności reszt aminokwasowych dla rozpuszczalnika (Faraggi et al., 2012),
- PSIPRED – program oparty na algorytmie ANN i profilach z programu PSI-BLAST, jeden z najczęściej stosowanych programów do przewidywania struktury drugorzędowej, bardzo często integrowany z innymi programami np. programami do przewidywania zwoju białka (McGuffin et al., 2000),
- SOPRANO – program przewiduje jednocześnie strukturę drugorzędową, dostępność reszt aminokwasowych dla rozpuszczalnika i zwroty β za pomocą zawansowanej wersji algorytmu ANN zwanej MOLEBRNN (Kirschner i Frishman, 2008),
- SSPAL – komercyjny serwis internetowy, używa lokalnego przyrównania sekwencji (Softberry, Inc.),
- SSP – komercyjny serwis internetowy, segmenty odpowiadające helisom α i wstęgom β przewidywane są za pomocą liniowej analizy dyskryminacyjnej (Softberry, Inc.),
- PSSFINDER – komercyjny serwis internetowy, przewiduje strukturę drugorzędową za pomocą łańcuchów Markova (Softberry, Inc.),
- NNSSP – komercyjny serwis internetowy (Softberry, Inc.), używa MSA i algorytmu najbliższego sąsiada (ang. *nearest-neighbor*), (Salamov i Solovyev, 1995),

- SSPRED – komercyjny serwis internetowy (Softberry, Inc.), algorytm ANN oparty na wyniku programu PSI-BLAST.

5.3.4.2. Programy do przewidywania domen

Pierwotnie wykrywanie domen na metaserwerze GeneSilico ograniczało się jedynie do przeszukiwania bazy PFAM za pomocą programu HMMER. Jednak obecnie istnieją także programy, które poza prostym przeszukiwaniem baz domen białkowych i identyfikacją domen na zasadzie podobieństwa przewidują domeny *ab initio*. W obrębie tej klasy możemy wydzielić dwa typy programów. Takie, które przewidują właściwe domeny (czyli podają ich liczbę, długość i lokalizacje) oraz programy, których działanie ograniczone jest do przewidywania miejsc stanowiących granice między sąsiadującymi ze sobą domenami. Ostatecznie wyniki wszystkich programów można sprowadzić do prostej formy tekstowej, gdzie poszczególnym domenom przyznawane są kolejne numery np. `---111...11111---222...222----...`. W przypadku metod przewidujących jedynie granice domen, poszczególne domeny nie są oddzielone od siebie łącznikami, chyba że metoda zwraca długość granicy. W chwili obecnej metaserwer umożliwia uruchomienie następujących programów:

- GLOBPLOT2 – program przewiduje globularne domeny na zasadzie statystyki prostych właściwości biochemicznych, przewidywania programu porównano do baz PFAM i SMART (Linding et al., 2003b),
- DOMpro – program oparty na algorytmie ANN działający na podstawie profili, struktury drugorzędowej oraz dostępności reszt aminokwasowych dla rozpuszczalnika. Wytrenowany na podstawie domen z bazy CATH (Cheng et al., 2006b),
- DOMAC – metoda ta składa się z dwóch etapów, najpierw identyfikowane są sekwencje homologiczne za pomocą programu PSI-BLAST, z nich tworzone są modele za pomocą programu MODELLER (Sali et al., 1995). Następnie domeny w strukturze modelu wykrywane są za pomocą programu PDP (Alexandrov i Shindyalov, 2003). Jeśli nie uda się znaleźć żadnych wiarygodnych homologów, sekwencja wysyłana jest do programu DOMpro (Cheng, 2007),
- Scooby-domain – program opiera swoje działanie na dwóch cechach: długość sekwencji i hydrofobowość, istnieje możliwość uruchomienia trybu, który dodatkowo uwzględnia wynik programu DomCut (Suyama i Ohara, 2003), czyli przewidywanie regionów

łączących domeny (pętle, zwroty β). Jako źródła danych użyto domen z bazy CATH (George et al., 2005),

- PPRODO – granice domen przewidywane są przez algorytm ANN oparty na macierzach PSSM (Sim et al., 2005),
- DomPred – program składa się z dwóch części: dla sekwencji dla których brak homologów o znanej strukturze uruchamiany jest program DGS (Wheelan et al., 2000), który opiera swoje działanie na długości sekwencji, natomiast w przypadku wysokiego podobieństwa do znanych sekwencji używany jest moduł o nazwie DomSSEA, który opiera się na przewidywaniu struktury drugorzędowej według programu PSIPRED. Program był trenowany na domenach z bazy CATH (Marsden et al., 2002),
- FIEFDom – program przewiduje granice domen używając logiki rozmytej, podstawą do podjęcia decyzji jest macierz PSSM porównana do biblioteki fragmentów zbudowanej z bazy domen SCOP (Bondugula et al., 2009),
- Shandy – serwis internetowy przewidujący granice domen przy użyciu rekurencyjnej sieci neuronowej. Cechy, które brane są pod uwagę to przewidywana struktura drugorzędowa, dostępność reszt aminokwasowych dla rozpuszczalnika, mapy kontaktów oraz szablony wraz z układem domen wziętym z bazy SCOP (Walsh et al., 2009),
- DoBo – program ten oparty jest na metodzie uczenia maszynowego SVM i uwzględnia macierz PSSM z programu PSI-BLAST, przewidywanie struktury drugorzędowej według programu SSpro oraz cechy takie jak długość sekwencji oraz pozycja aminokwasu względem jej końców (Eickholt et al., 2011),
- DomainSVM – program przewiduje występowanie domen opierając się na podstawie cech takich jak entropia, hydrofobowość, przewidywana struktura drugorzędowa, dostępność reszt aminokwasowych dla rozpuszczalnika, występowanie regionów wewnętrznie nieuporządkowanych oraz obecność bliskich homologów w bazach domen białkowych CATH i PFAM. Opiera on swoje działanie na metodzie SVM (program ten jest jednym z głównych wyników niniejszej rozprawy doktorskiej, jego szczegółowy opis znajduje się w dalszej części rozprawy).

5.3.4.3. Programy do przewidywania regionów wewnętrznie nieuporządkowanych

Tuż po odkryciu białek IUP, gdy zdano sobie sprawę z ich znaczenia, zaczęły powstawać kolejne, coraz bardziej dokładne programy do przewidywania regionów wewnętrznie nieuporządkowanych. W chwili obecnej dostępnych jest ponad 60 programów tego typu (szczegółowa lista zamieszczona została na stronie internetowej http://iimcb.genesilico.pl/metadisorder/list_of_protein_disorder_tools_programs.html).

Nie wszystkie te programy spełniają jednak warunki niezbędne do integracji z metaserwerem GeneSilico. Część programów przestaje być dostępna w pewnym momencie, inne nie są w stanie obsłużyć dużej liczby zapytań jakie wysyła metaserwer, czasem licencja programu nie pozwala na włączanie go do metaserwerów lub program dostępny jest jedynie w wersji płatnej. Głównymi kryteriami wyboru programów, które są zintegrowane z metaserwerem są: niezawodność (m.in. dlatego lokalne programy są preferowane w stosunku do serwisów internetowych) oraz względna jakość metody (np. w oparciu o wynik w konkursie CASP). Wynik poszczególnych metod prezentowany jest w formie ciągu znaków: `DDDDD-----DD...`, gdzie „D" oznacza resztę w obrębie regionu wewnętrznie nieuporządkowanego, a „−" oznacza ustrukturalizowaną część białka. Oprócz wyników poszczególnych metod prezentowany jest także konsensus. W chwili obecnej metaserwer umożliwia uruchomienie następujących programów:

- DisEMBL – program ten pozwala przewidywać regiony IUR zgodnie z jedną z trzech definicji nieustrukturalizowania (REMARK465, regiony pozbawione struktury drugorzędowej oraz regiony o wysokiej wartości czynnika temperaturowego), program wytrenowano w oparciu na białkach z bazy danych SCOP, do budowy klasyfikatora wykorzystano algorytm ANN (Linding et al., 2003a),
- GLOBPLOT – metoda opiera się na statystyce występowania struktury drugorzędowej i pętli dla reprezentatywnej próbki białek z bazy SCOP (Linding et al., 2003b),
- DISOPRED2 – przewiduje regiony IUR zdefiniowane są jako brakujące atomy w strukturach z bazy PDB (REMARK465), wykorzystano algorytm SVM w oparciu na profilach PSSM i przewidywaną strukturę drugorzędową (Ward et al., 2004),
- DISPROT (VSL2B) – grupa metod, które oparte są na algorytmie SVM i kombinacji różnych właściwości sekwencyjnych takich jak częstotliwość aminokwasów w macierzy PSSM, hydrofobowość, ładunek, entropia oraz przewidywana struktura drugorzędowa.

Jako zbiór uczący wykorzystano białka z PDB oraz z bazy DISPROT (Obradovic et al., 2005),

- IUPred – program pozwala przewidywać długie i krótkie regiony IUR na podstawie energii odziaływania między resztami aminokwasowymi (Dosztanyi et al., 2005),
- DISpro – algorytm ANN wykorzystujący profil PSSM oraz przewidywania struktury drugorzędowej i dostępności reszt aminokwasowych dla rozpuszczalnika (Cheng et al., 2005b),
- RONN – program implementuje algorytm ANN działający na podstawie prawdopodobieństwa występowania stanu nieuporządkowania w przyrównaniu wielu sekwencji (Yang et al., 2005),
- SPRITZ – serwis internetowy przewidujący regiony IUR oparty na algorytmie SVM oraz profilu PSSM, przewidywaniu struktury drugorzędowej i dostępności reszt aminokwasowych dla rozpuszczalnika, posiada dwa tryby pozwalające na przewidywanie długich i krótkich regionów nieuporządkowanych (Vullo et al., 2006),
- PDISORDER – komercyjny serwis internetowy, nieuporządkowanie przewidywane jest przez algorytm ANN i liniową analizę dyskryminacyjną (Softberry, Inc.),
- POODLE-L – serwis internetowy oparty na dwupoziomowym algorytmie SVM oraz cechach takich jak hydrofobowość, entropia Shannona, ładunek, przewidywana struktura drugorzędowa i kontakty między resztami aminokwasowymi (Hirose et al., 2007),
- POODLE-S – serwis internetowy przewidujący krótkie fragmenty IUR na podstawie profili PSSM właściwości fizykochemicznych takich jak hydrofobowość, ładunek, wielkość i polarność (Shimizu et al., 2007),
- PrDOS – serwis internetowy opierający się na algorytmie SVM, w którym dane wejściowe stanowi macierz PSSM oraz dodatkowo brana jest pod uwagę informacja strukturalna pochodząca ze struktur homologicznych (Ishida i Kinoshita, 2007),
- iPDA – serwer internetowy o rozszerzonej funkcjonalności w stosunku do swojego poprzednika – programu DisPSSMP2 (Su et al., 2006) o przewidywania hydrofobowości, konserwacji aminokwasowej, detekcję regionów o niskiej złożoności oraz przewidywanie struktury drugorzędowej; opiera się na specjalnym rodzaju algorytmu ANN, tzw. sieci radialnej (ang. *radial basis function network*) (Su et al., 2007),
- Metadisorder – meta-metoda oparta na algorytmie ANN, uwzględnia strukturę

drugorzędową, dostępności reszt aminokwasowych dla rozpuszczalnika, przewidywanie niskiej złożoności sekwencyjnej oraz profil PSSM (Schlessinger et al., 2009),

- SPINE-D – algorytm ANN w którym dane wejściowe stanowią: właściwości fizykochemiczne aminokwasu, macierz PSSM, przewidywana struktura drugorzędowa oraz kąty torsyjne (Zhang et al., 2012),
- GeneSilico MetaDisorder – grupa metaprogramów opartych na wynikach z innych metod przewidujących regiony IUR, strukturze drugorzędowej oraz pokryciu przyrównania wielu sekwencji dla szablonów wykrytych przez metody do rozpoznawania zwoju. Konsensus powstaje przy wykorzystaniu algorytm genetycznego, program ten jest jednym z głównych wyników niniejszej rozprawy doktorskiej, jego szczegółowy opis znajduje się w dalszej części rozprawy (Kozlowski i Bujnicki, 2012).

5.3.4.4. Programy do przewidywania dostępności reszt aminokwasowych dla rozpuszczalnika

Przewidywanie dostępności reszt aminokwasowych dla rozpuszczalnika jest ważnym elementem przewidywania struktury, ponieważ pozwala określić lokalizację reszty w stosunku do powierzchni białka. Większość metod przewiduje relatywną dostępność dla rozpuszczalnika (RSA) i stosując odpowiedni próg (zwykle stosuje się progi 25%, 5% i 0%) określa, które reszty nie są dostępne dla rozpuszczalnika. Wynik poszczególnych metod prezentowany jest w formie ciągu znaków: `---BBBB-BB-BBB`…, gdzie „B” oznacza resztę zagrzebaną, niedostępną dla rozpuszczalnika, a „-” oznacza resztę dostępną dla rozpuszczalnika przy określonym progu. Dostępne programy to:

- NetSurfP – wielowarstwowy algorytm ANN oparty na profilach z program PSI-BLAST oraz przewidywanej strukturze drugorzędowej, program przewiduje zagrzebanie na poziomie 25% (Petersen et al., 2009),
- SOPRANO – program przewiduje dostępność reszt aminokwasowych dla rozpuszczalnika na poziomie 25% oraz dodatkowo strukturę drugorzędową i zwroty β za pomocą zaawansowanej wersji algorytmu ANN zwanej MOLEBRNN (Kirschner i Frishman, 2008),
- ACCpro – program opiera swoje działanie na profilach PSSM otrzymane za pomocą programu PSI-BLAST, które są przetwarzane przez sztuczną sieć neuronową. Możliwe

jest ustawienie dowolnego progu dostępności reszt z dokładnością do 5% (Pollastri et al., 2002),

- JNET – algorytm ANN oparty na przyrównaniu wielu sekwencji, program przewiduje dostępność reszt aminokwasowych dla rozpuszczalnika na poziomie 25%, 5% i 0% (Cole et al., 2008),
- SPINE – wykorzystuje algorytm ANN, który obok profili otrzymanych za pomocą program PSI-BLAST jako dane wejściowe bierze pod uwagę, także takie cechy jak hydrofobowość, objętość, polarność oraz punkt izoelektryczny (Dor i Zhou, 2007),
- SPINE X – program opiera swoje działanie na profilach PSSM, przewidywaną strukturę drugorzędową oraz kąty torsyjne, które są przetwarzane przez algorytm ANN (Faraggi et al., 2012),
- SABLE – wartości RSA oparte są na modelu regresji liniowej (Wagner et al., 2005),
- WESA – metaprogram oparty na wynikach z drzewa decyzyjnego, statystyki Bayesa, wielokrotnej regresji linowej oraz algorytmów ANN i SVM, (Tjong et al., 2007).

5.3.4.5. Programy do przewidywania helis transbłonowych

Pierwotnie metaserwer zawierał trzy programy do przewidywania helis transbłonowych: MEMSAT2 (Jones et al., 1994), TMHMM (Sonnhammer et al., 1998), TMPred (Hofmann i Stoffel, 1993). Wraz z nagromadzeniem wiedzy powstawały kolejne, lepsze programy, które zostały zintegrowane z metaserwerem. Najnowsze programy są w stanie nie tylko określić lokalizację helis transbłonowych, ale także potrafią ocenić kierunek poszczególnych pętli (lokalizacja na zewnątrz lub wewnątrz komórki) oraz odróżnić je od sekwencji sygnałowych. Wyniki poszczególnych metod prezentowane są w formie ciągu znaków np.:

SSSSSSSSSSSSSS----------HHHHHHHHHHH++++++++++HHH...

gdzie „S” oznacza sekwencję sygnalną zlokalizowaną na końcu N, „-” oznacza pętle skierowaną na zewnątrz komórki, „+” oznacza pętle skierowaną do wnętrza, a „H” oznacza region helisy transbłonowej. W chwili obecnej metaserwer umożliwia uruchomienie następujących programów:

- Phobius – program opiera swoje działanie na modelach HMM wzbogaconych o informację dostępną dla najbliższych homologów. Program dodatkowo potrafi przewidzieć sekwencję sygnalną na końcu N (Kall et al., 2007),

- MEMSAT3 – używa algorytmu ANN, który wykorzystuje macierze PSSM z programu PSI-BLAST. W stosunku do poprzedniej wersji program dodatkowo odróżnia sekwencje sygnalne od helis transbłonowych (Jones, 2007),
- TMPred – program opiera swoje działanie na analizie statystycznej profili z bazy białek transbłonowych TMbase (Hofmann i Stoffel, 1993),
- HMMTOP – program używa modeli HMM, jego cechą szczególną jest to, że pozwala na wprowadzenie więzów wynikających z informacji na temat lokalizacji znanych motywów (Tusnady i Simon, 2001),
- DAS – program porównuje profil hydrofobowości do profili znanych białek transbłonowych (Cserzo et al., 2004),
- MINNOU – program wykorzystuje algorytm ANN oraz profile PSSM, hydrofobowość, przewidywaną strukturę drugorzędową oraz dostępność reszt aminokwasowych dla rozpuszczalnika (Cao et al., 2006),
- TMHMM2.0 – program opiera swoje działanie na modelach HMM, ponadto pozwala przewidywać sekwencję sygnalną (Krogh et al., 2001),
- PRODIV, PRO, S_TMHMM – pierwsze dwa programy wykorzystują profile PSSM w modelu HMM, natomiast S_TMHMM jest wersją programu TMHMM przetrenowaną ponownie na nowym zbiorze testowym wspólnym dla wszystkich trzech programów (Viklund i Elofsson, 2004),
- SCAMPI-seq, SCAMPI-msa – przewidywanie topologii helis transbłonowych wykorzystujące skalę aminokwasową obrazującą zależność zmiany energii swobodnej ΔG w czasie integracji białka w błonę biologiczną, dodatkowo druga wersja programu wykorzystuje przyrównanie wielu sekwencji (Bernsel et al., 2008),
- OCTOPUS – przewidywanie helis transbłonowych odbywa się przez kombinację modeli HMM z algorytmem ANN, program rozróżnia helisy transbłonowe od sekwencji sygnalnych (Viklund i Elofsson, 2008),
- TOPCONS – metaprogram, którego wynikiem jest konsensus z programów: OCTOPUS, PRO, PRODIV, SCAMPI-seq oraz SCAMPI-msa (Bernsel et al., 2009),
- Proteus – serwis internetowy, który do przewidywania helis transbłonowych używa programu TMHMM oraz dodatkowo sprawdza podobieństwo do znanych sekwencji sygnałowych i białek cytoplazmatycznych (Montgomerie et al., 2008).

5.3.4.6. Programy do przewidywania struktur splecionych helis

Metaserwer umożliwia uruchomienie następujących programów do przewidywania struktur splecionych helis:

- COILS – jeden z pierwszych programów do przewidywania struktur splecionych helis, używa informacji na temat częstości występowania określonych reszt aminokwasowych w motywach splecionych helis (Lupas et al., 1991),
- PAIRCOIL – program opiera swoje działanie na macierzy PSSM (Berger et al., 1995),
- MARCOIL – program stanowi rozszerzenie programu PAIRCOIL, potrafi dodatkowo przewidywać liczbę helis tworzących strukturę superhelisy (Wolf et al., 1997),
- PCOILS – opiera swoje działanie o macierze PSSM z programu PSI-BLAST (Gruber et al., 2005),
- CAST – program przewiduje motywy splecionych helis biorąc pod uwagę niską złożoność sekwencyjna takich regionów (Promponas et al., 2000),
- NCOILS – opiera swoje działanie o maszynę regresji wektorów nośnych (ang. *support vector regression*), program autorstwa Roba Russella (dane nieopublikowane).

5.3.4.7. Programy do przewidywania oddziaływania białek z DNA i białek z RNA

Pod pojęciem oddziaływania białek z DNA i białek z RNA rozumiemy możliwość tworzenia oddziaływania między fragmentem białka a kwasem nukleinowym. Oznacza to, że w strukturze kompleksu reszty aminokwasowe znajdują się w określonej odległości od cząsteczki RNA lub DNA, zwykle jako graniczną przyjmuje się odległość wynoszącą 3,5 Å lub 5 Å. Wyniki poszczególnych programów z tej kategorii są prezentowane w formie ciągu znaków np. `---++++++---++-?--...`, gdzie „+” oznacza resztę aminokwasową oddziałującą z DNA lub RNA,

„-” oznacza brak oddziaływania, natomiast „?” oznacza, że program nie był w stanie wiarygodnie ocenić danej cechy. Na podstawie otrzymanych wyników liczony jest prosty konsensus w postaci średniej arytmetycznej. Ponadto dodany został także konsensus (oznaczony jako cons3best) wzbogacony o system wag odpowiadających jakości metody (Puton et al., 2012). Programy włączone do metaserwera GeneSilico to:

- DP-BIND – serwis internetowy przewidujący wiązanie białek z DNA na podstawie

konsensusu z trzech metod uczenia maszynowego: metody SVM oraz dwóch rodzajów regresji liniowej. Jako dane wejściowe programy te używają macierzy PSSM, struktury drugorzędowej i dostępności reszt aminokwasowych dla rozpuszczalnika (Hwang et al., 2007),

- DISIS – program do identyfikacji reszt wiążących DNA na podstawie ich sąsiedztwa w sekwencji, profilu ewolucyjnego, przewidywanej dostępności dla rozpuszczalnika i struktury drugorzędowej białka, metoda oparta na algorytmie SVM (Ofran et al., 2007),
- BindN – serwis internetowy przewidujący odziaływanie białka z DNA na podstawie masy atomowej, hydrofobowości i wartości pK dla reszt aminokwasowych, oparty na algorytmie SVM (Wang i Brown, 2006),
- BindN-RF – przewiduje oddziaływanie białek z DNA przy wykorzystaniu techniki lasów losowych (ang. *random forest*), prócz właściwości, których użyto w programie BindN, dodatkowo wykorzystuje informację ewolucyjną w postaci macierzy PSSM (Wang et al., 2009),
- BindN+ – serwis internetowy, który pozwala przewidywać zarówno odziaływania białek z RNA, jaki i białko DNA, opiera się na algorytmie SVM i tych samych właściwościach co program BindN-RF (Wang et al., 2010),
- DBindR – serwis internetowy służący do przewidywania oddziaływania białek z DNA, opiera swoje działanie na metodzie lasów losowych oraz metodzie SVM, wykorzystuje takie cechy jak: profile PSSM, struktura drugorzędowa, typ reszty aminokwasowej (Wu et al., 2009),
- NAPS – serwis internetowy, który na podstawie rodzaju reszty aminokwasowej, jej ładunku oraz konserwacji reszt położonych w jej sąsiedztwie przewiduje wiązanie białek z DNA/RNA korzystając z drzewa decyzyjnego C4.5 (Carson et al., 2010),
- PPRInt – serwis internetowy identyfikujący odziaływanie białek z RNA oparty na algorytmie SVM i profilach PSSM (Kumar et al., 2007),
- PiRaNhA – serwis internetowy służący do przewidywania oddziaływania białek z DNA oparty na algorytmie SVM używający macierzy PSSM, częstości występowania poszczególnych reszt aminokwasowych w miejscach oddziaływania, przewidywanej dostępności reszt aminokwasowych dla rozpuszczalnika oraz hydrofobowości (Murakami et al., 2010),

- RNABindR – serwis internetowy do przewidywania odziaływania białek z RNA, działanie programu opiera się na wykorzystaniu naiwnego klasyfikatora Bayesa (Terribilini et al., 2007),
- cons3best – meta-metoda do przewidywania oddziaływania białek z RNA oparta na oparta na wynikach z programów PPRInt, PiRaNhA i BindN+ (Puton et al., 2012).

5.3.4.8. Programy do przewidywania mostków dwusiarczkowych

Programy przewidujące występowanie wiązań dwusiarczkowych między cysteinami dostępne w chwili obecnej na metaserwerze to:

- CYS_REC – komercyjny serwis internetowy oparty na algorytmie ANN (Softberry, Inc.),
- DiANNA – serwis internetowy, który wykorzystuje technikę SVM w celu przewidywania wiązań dwusiarczkowych oraz oddziaływania cysteiny z ligandami (np. żelazem, cynkiem, kadmem) (Ferre i Clote, 2006),
- DIpro2 – program używa rekurencyjnego algorytmu ANN, który wykorzystuje profile PSSM, strukturę drugorzędową i dostępność reszt aminokwasowych dla rozpuszczalnika wziętych z plików PDB najbliższych homologów lub/i przewidzianych za pomocą programu DSSP (Cheng et al., 2006a),
- DISULFIND – algorytm, który składa się z dwóch etapów, najpierw metoda SVM ocenia czy dana cysteina uczestniczy w tworzeniu wiązania dwusiarczkowego, a następnie dla pozytywnych przypadków za pomocą algorytmu ANN przewidywany jest wzór wiązania się cystein (Ceroni et al., 2006),
- DBCP – serwis internetowy, którego działanie oparte jest na metodzie SVM. Jako dane wejściowe wykorzystywane są współrzędne wszystkich cystein w modelu wygenerowanym przez program MODELLER (Lin i Tseng, 2010).

5.3.4.9. Programy do wykrywania zwoju białka

Programy do rozpoznawania zwoju (ang. *fold recognition*) stanowią główna kategorię programów wchodzących w skład metaserwera GeneSilico. Ich zadaniem jest wykrycie najlepszych szablonów i przez to umożliwienie zbudowania jak najlepszych modeli homologicznych. Schemat działania programów tego typu jest następujący: najpierw za pomocą

sekwencji-celu w dużej bazie sekwencyjnej np. nr lub uniprot wyszukiwane są sekwencje homologiczne, a następnie otrzymane przyrównanie MSA wykorzystywane jest do znalezienia homologów w bazie struktur np. PDB. Różnice między metodami dotyczą sposobu reprezentacji przyrównania MSA, budowania macierzy PSSM czy modelu HMM na jego podstawie oraz bazach danych użytych do przeszukiwania. Wyniki poszczególnych metod prezentowane są w postaci oddzielnych tabel zawierających do dziesięciu najlepszych zidentyfikowanych szablonów. Przy każdej sekwencji szablonu podany jest kod PDB, poziom istotności statystycznej, procent identyczności sekwencji-celu i szablonu, numer rodziny SCOP (uzyskany z pliku PDB lub przewidziany za pomocą programu fastSCOP (Tung i Yang, 2007)), numer EC reprezentujący klasyfikację enzymatyczną (uzyskany z pliku PDB lub przewidziany za pomocą programu EnzyPred (Shen i Chou, 2007)). Dodatkowo sekwencja szablonu pokolorowana jest zgodnie z jego strukturą drugorzędową (czerwony-helisy, zielony-wstęgi). Ponadto dla każdego z szablonów możliwe jest pobranie przyrównania szablonu i celu w formacie dostosowanym do programu MODELLER oraz prostego modelu opartego na współrzędnych z pliku PDB szablonu. W chwili obecnej metaserwer umożliwia uruchomienie następujących programów:

- PSI-BLAST – wersja programu BLAST, która używa strategii iteracyjnego przeszukiwania bazy danych poprzez tworzenie lokalnego przyrównania wielu sekwencji, a następnie przekształcanie go w profil, który jest wykorzystywany do wyszukiwania kolejnych sekwencji dodawanych do przyrównania program BLAST stanowi heurystyczną aproksymację algorytmu Smith-Watermana nie gwarantującą znalezienia najlepszego przyrównania sekwencji, ale jest ponad 50 razy szybszy od programu FASTA, Metaserwer uruchamia program PSI-BLAST na bazie sekwencji z PDB używając dwóch oddzielnych ustawień (z włączonym i wyłączonym filtrowaniem regionów o niskiej złożoności sekwencyjnej, oznaczone odpowiednio jako „pdblast” i „blastp”) (Altschul et al., 1997),
- GenTHREADER – metoda porównująca sekwencje celu do bazy profili, cechą szczególną jest wykorzystanie algorytmu ANN do oceny istotności podobieństwa sekwencji do profilu (Jones, 1999),
- COMPASS – wybór szablonów opiera się na analitycznym przybliżeniu wartości E (ang. *E-value*) dla otrzymanych przyrównań MSA (Sadreyev et al., 2007),
- PRC – program opiera się na porównaniu profilu HMM badanego białka do biblioteki

profili HMM białek o znanych strukturach, dodatkowo jako dane wejściowe mogą być użyte pliki wynikowe z innych programów takich jak SAM (Karplus, 2009), HMMER oraz PSI-BLAST (Madera, 2008),

- COMA – ang. ***C**omparison **O**f **M**ultiple **A**lignments* – metoda porównująca profile z profilami, z sekwencji wejściowej generowany jest profil, który porównywany jest z bazą profili (Margelevicius i Venclovas, 2010),
- CS-BLAST – ang. *context specific BLAST* – wariant programu BLAST, który poprawę działania w stosunku do podstawowej wersji programu BLAST zawdzięcza wykorzystaniu 12 aminokwasowego okna, które reprezentuje otoczenie reszty aminokwasowej (Biegert i Soding, 2009),
- HHblits – program opiera się na porównaniu profilu HMM do biblioteki profili HMM, metoda ta jest szybsza i dokładniejsza niż PSI-BLAST (Remmert et al., 2012),
- HHsearch – metoda profil-profil, w której przyrównanie MSA jest przedstawione w formie modelu HMM, którym przeszukiwana jest biblioteka modeli HMM. Metaserwer wykorzystuje bazy CDD i PDB70; najnowszej wersji program nie korzysta już z programu PSI-BLAST do budowania pierwotnego przyrównania, lecz wykorzystuje program HHblits (Soding, 2005),
- FFAS – metoda profil-profil, jej cechą szczególną jest sposób budowania profilu PSSM, w którym wagi sekwencji z przyrównania MSA odpowiadają poziomowi zgodności z sekwencją konsensusową reprezentującą daną rodzinę białkową (Rychlewski et al., 2000),
- Sp3 – metoda profil-profil, dodatkowo prócz standardowej procedury przeszukiwania bazy profili przeprowadzane jest także porównanie sekwencji do biblioteki fragmentów strukturalnych, ponadto brany jest także pod uwagę wynik przewidywania struktury drugorzędowej (Zhou i Zhou, 2005),
- Phyre – metaserwer, który w czasie porównywania profilu do bazy profili za pomocą programu HHserach wykorzystuje dodatkowo takie cechy jak: przewidywana struktura drugorzędowa (PSIPRED), brak uporządkowania struktury trzeciorzędowej (DISOPRED) oraz obecność helis transbłonowych (MEMSAT) (Bennett-Lovsey et al., 2007),
- HMMER – jeden z powszechniej wykorzystywanych programów typu profil-profil, w

metaserwerze odpowiada za identyfikację rodzin białkowych z bazy PFAM (Finn et al., 2011),

- PCONS – metoda, która ocenia jakość modeli za pomocą algorytmu ANN. Pod uwagę brane są zgodność struktury drugorzędowej oraz jakość modelu według programu ProQ (Wallner i Elofsson, 2003) służącego do oceny ogólnej jakości modeli (MQAP, ang. *model quality assessment program*), modele będące danymi wejściowymi dla programu PCONS budowane są automatycznie w oparciu na przyrównaniach do szablonów wykrytych przez wcześniej wymienione metody do rozpoznawania zwoju (Wallner i Elofsson, 2005).

5.3.5. Programy i serwisy internetowe stanowiące część serwisu internetowego do przewidywania regionów wewnętrznie nieuporządkowanych

Serwis internetowy do przewidywania regionów wewnętrznie nieuporządkowanych opiera swoje działanie na wynikach programów, które można podzielić na kilka kategorii. Pierwsza i najważniejsza grupa to niezależnie opracowane przez inne grupy badawcze programy do przewidywania braku uporządkowania: DisEMBL, GLOBPLOT, DISOPRED2, DISPROT (VSL2B), IUPred (uruchamiany w dwóch trybach, dostosowanych odpowiednio do przewidywania długich i krótkich regionów IUR), DISpro, RONN, SPRITZ, PDISORDER, POODLE-L, POODLE-S, PrDOS oraz iPDA. Następna grupa programów wykorzystana w opisywanym serwisie to programy do wykrywania zwoju: PSI-BLAST (z włączonym i wyłączonym filtrowaniem regionów o niskiej złożoności sekwencyjnej, oznaczone odpowiednio jako pdblast i blastp), FFAS, HHsearch (na bazach pdb70 i cdd), GenTHREADER, Phyre i PCONS. Ponadto do przewidywania struktury drugorzędowej wykorzystano program SSPro4.

5.3.6. Programy stanowiące część serwisu internetowego do przewidywania domen białkowych

W przeciwieństwie do programu przewidującego regiony wewnętrznie nieuporządkowane, metoda służąca do przewidywania domen opisana w ramach niniejszej rozprawy nie jest oparta na standardowej metodologii tworzenia meta-przewidywań. Serwis nie korzysta z wyników innych metod do przewidywania domen. Wykorzystane zostały jedynie

programy do przewidywania innych cech białka takich jak struktura drugorzędowa (SSPro4 i PSIPRED), brak uporządkowania (RONN) oraz homologia do białek o znanych strukturach (HHblits).

5.3.7. Programy wykorzystane w trakcie analizy kompleksu odpowiedzialnego za obróbkę końca 3′ mRNA

Podstawowym narzędziem wykorzystanym w trakcie analizy kompleksu odpowiedzialnego za obróbkę końca 3′ mRNA był metaserwer GeneSilico. Za jego pomocą przeprowadzono poszukiwania homologii do znanych struktur białkowych oraz wygenerowano modele homologiczne. Regiony wewnętrznie nieuporządkowane przewidziano za pomocą programu GeneSilico MetaDisorder. Domeny białkowe oznaczono według następującego schematu: najpierw pobrano dostępne dane z bazy domen PFAM, a następnie porównano je z granicami domen przewidzianymi przez program DomainSVM. Regiony o niskiej złożoności sekwencyjnej przewidziano za pomocą programu SEG (Wootton, 1994).

6. Wyniki

6.1. Metaserwer GeneSilico

Metaserwer GeneSilico dostępny jest jako serwis internetowy pod adresem https://www.genesilico.pl/meta2. Pierwotnym zadaniem serwisu było ułatwienie homologicznego modelowania białek. Główną jego częścią były programy do rozpoznawania zwoju, które służyły do zidentyfikowania najlepszego szablonu do modelowania. Początkowo, sumaryczna liczba programów wchodzących w skład metaserwera nie przekraczała 20. W chwili obecnej metaserwer pozwala w łatwy sposób uruchomić ponad 120 programów bioinformatycznych, które przewidują nie tylko zwój, ale również takie cechy białka jak struktura drugorzędowa, obecność helis transbłonowych, sekwencji sygnalnych, mostków dwusiarczkowych i struktur splecionych helis, dostępność reszt aminokwasowych dla rozpuszczalnika, domeny oraz wewnętrzne nieuporządkowanie (tabela 2). Programy, o które rozbudowano metaserwer zostały zainstalowane przez autora rozprawy. Każdy z nich wymagał napisania oddzielnego skryptu uruchamiającego i przetwarzającego jego wynik do jednorodnego formatu. Skrypty te musiały uwzględniać specyfikę programu (format danych wejściowych, konfigurację sprzętu, zmienne środowiskowe, itp.). Ponadto, większość skryptów pochodzących z pierwotnej wersji serwisu także musiała być przepisana, ponieważ wraz z wzrostem liczby programów zastosowane rozwiązania okazały się niewystarczające. Autorem wspomnianych skryptów jest autor rozprawy.

Przy wyborze programów, które mają być włączone do metaserwera preferowane były te, które można zainstalować lokalnie, ponieważ raz zainstalowane nie przysparzają już zwykle problemów. Uruchamianie i przetwarzanie wyników w zautomatyzowany sposób jest względnie łatwe, a lista potencjalnych błędów koniecznych do wychwycenia jest stosunkowo krótka. Na tym tle zewnętrzne serwisy internetowe, które również w miarę możliwości włączane są do metaserwera, prezentują się jako wymagające większego wysiłku. Skrypty zarządzające tego typu metodami muszą być w stanie poradzić sobie z niestandardowymi sytuacjami takimi jak chwilowe wyłączenie serwisu zewnętrznego, konieczność ponownego wysyłania zapytania w przypadku przekroczenia czasu odpowiedzi, pobieranie wiadomości ze skrzynki email i z wieloma innymi trudnymi do przewidzenia sytuacjami.

Oddzielną grupę programów stanowią programy, które pełnią funkcję pomocnicze i służą do lepszej prezentacji danych (np. program DSSP (Kabsch i Sander, 1983) wykorzystany do kolorowania sekwencji według struktury drugorzędowej szablonu). Dodatkowo z metaserwerem połączone są także inne programy, które mogą generować dalsze wyniki na podstawie bezpośrednio pobranych z metaserwera danych (np. program AmIgoMR (Pawlowski i Bujnicki, 2012), który korzysta z modeli generowanych przez metaserwer do przygotowania plików wejściowych do krystalograficznej procedury podstawienia molekularnego).

Wzrost liczby programów wchodzących w skład metaserwera oraz ograniczone zasoby komputerowe powodowały, że niezbędne było wprowadzenie pewnych ograniczeń możliwości użytkownika. Domyślnie użytkownik może wysłać maksymalnie 20 sekwencji na dzień i nie może wysyłać jako zapytanie takiej samej sekwencji ponownie. Ponadto długość sekwencji ograniczona jest do zakresu od 40 do 990 aminokwasów. Ostatnie wymaganie wynika z tego, że dla sekwencji krótszych niż 40 aminokwasów większość metod daje mało wiarygodne wyniki lub nie zwraca ich wcale. Podobnie jest w przypadku bardzo dużych białek zbudowanych z tysiąca i więcej reszt aminokwasowych. Są one prawie zawsze zbudowane z kilku domen, a większość metod trenowana była na krótkich białkach, zwykle jednodomenowych. Należy, więc oczekiwać, że dla białek zbudowanych z wielu domen ich wyniki będą znacząco gorsze. Poza tym domeny białkowe można traktować jako niezależne jednostki. Ich proces zwijania jest zwykle jedynie w ograniczonym stopniu zależny od innych domen w białku. W związku z tym zalecane jest uruchamianie większości narzędzi bioinformatycznych dla fragmentów sekwencji odpowiadających pojedynczym domenom lub grupom domen tworzących zwartą jednostkę strukturalną.

Tabela 2. Programy bioinformatyczne wchodzące w skład metaserwera GeneSilico. Bardziej szczegółowy opis poszczególnych programów można znaleźć w rozdziale „Materiały i metody". W kolumnie „Kategoria" oznaczenie „zdalny szybki" oznacza serwis zewnętrzny, który zwraca wynik bezpośrednio po wysłaniu zapytania na oddzielnej stronie HTML, „zdalny wolny" oznacza serwis zewnętrzny, który odsyła wyniki na podany adres email, natomiast „lokalny" oznacza program zainstalowany lokalnie. Typ programu „inne" oznacza programy pomocnicze oraz programy zewnętrzne, które mogą być uruchamiane na podstawie wyników metaserwera. Kursywą zaznaczono programy, które były obecne w pierwotnej wersji metaserwera w momencie publikacji w 2003 roku (część z tych programów została zastąpionych przez nowsze wersje). Pogrubioną czcionką zaznaczono programy, których autorem jest autor niniejszej rozprawy. [1] programy wchodzące w skład programu MetaDisorder; [2] programy wchodzące w skład programu MetaDisorder3D, [3] programy wchodzące w skład programów MetaDisorderMD i MetaDisorderMD2, [4] programy wchodzące w skład programu DomainSVM.

Nazwa programu	Przewidywana cecha	Kategoria	Referencja
HMMSTR	struktura drugorzędowa	lokalny	(Bystroff et al., 2000)
SSPro4[3,4]	struktura drugorzędowa	lokalny	(Pollastri et al., 2002)
JNET	struktura drugorzędowa	lokalny	(Cole et al., 2008)
GOR	struktura drugorzędowa	zdalny wolny	(Kloczkowski et al., 2002)
FMD	struktura drugorzędowa	zdalny wolny	(Cheng et al., 2005a)
CDM	struktura drugorzędowa	zdalny wolny	(Cheng et al., 2007)
PROTEUS2	struktura drugorzędowa	zdalny wolny	(Montgomerie et al., 2008)
*SPARROW	struktura drugorzędowa	lokalny	(Bettella et al., 2012)
SPARROW	struktura drugorzędowa	lokalny	(Bettella et al., 2012)
SABLE	struktura drugorzędowa	lokalny	(Adamczak et al., 2005)
PROF PHD	*struktura drugorzędowa*	*lokalny*	*(Rost i Sander, 1993)*
PORTER	struktura drugorzędowa	zdalny wolny	(Pollastri i McLysaght, 2005)
NetSurfP	struktura drugorzędowa	lokalny	(Petersen et al., 2009)
RaptorX-SS8	struktura drugorzędowa	lokalny	(Wang et al., 2011)
PSSpred	struktura drugorzędowa	lokalny	(Zhang, 2012)
SPINE	struktura drugorzędowa	lokalny	(Dor i Zhou, 2007)
SPINE X	struktura drugorzędowa	lokalny	(Faraggi et al., 2012)
PSIPRED[4]	*struktura drugorzędowa*	*lokalny*	*(McGuffin et al., 2000)*
SOPRANO	struktura drugorzędowa	lokalny	(Kirschner i Frishman, 2008)
SSPAL	struktura drugorzędowa	zdalny szybki	Softberry, Inc.
SSP	struktura drugorzędowa	zdalny szybki	Softberry, Inc.
PSSFINDER	struktura drugorzędowa	zdalny szybki	Softberry, Inc.
NNSSP	struktura drugorzędowa	zdalny szybki	(Salamov i Solovyev, 1995)
SSPRED	struktura drugorzędowa	zdalny szybki	Softberry, Inc.
HMMER	domeny	lokalny	(Finn et al., 2011)
GLOBPLOT2	domeny	lokalny	(Linding et al., 2003b)
DOMpro	domeny	lokalny	(Cheng et al., 2006b)
DOMAC	domeny	zdalny wolny	(Cheng, 2007)
Scooby-domain	domeny	lokalny	(George et al., 2005)
PPRODO	domeny	lokalny	(Sim et al., 2005)
DomPred	domeny	lokalny	(Marsden et al., 2002)
FIEFDom	domeny	lokalny	(Bondugula et al., 2009)
Shandy	domeny	zdalny wolny	(Walsh et al., 2009)
DoBo	domeny	zdalny wolny	(Eickholt et al., 2011)
DomainSVM	**domeny**	**lokalny**	-
GLOBPLOT[1,3]	brak uporządkowania	lokalny	(Linding et al., 2003b)
DisEMBL[1,3]	brak uporządkowania	lokalny	(Linding et al., 2003a)

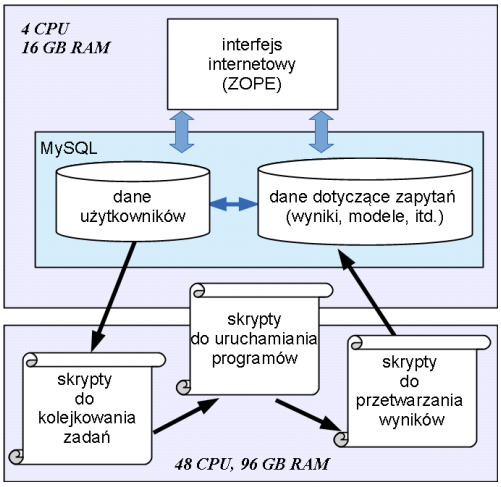
Nazwa programu	Przewidywana cecha	Kategoria	Referencja
DISOPRED2[1,3]	brak uporządkowania	lokalny	(Ward et al., 2004)
DISPROT (VSL2B) [1,3]	brak uporządkowania	lokalny	(Obradovic et al., 2005)
IUPred[1,3]	brak uporządkowania	lokalny	(Dosztanyi et al., 2005)
DISpro[1,3]	brak uporządkowania	lokalny	(Cheng et al., 2005b)
RONN[1,3,4]	brak uporządkowania	lokalny	(Yang et al., 2005)
SPRITZ[1,3]	brak uporządkowania	zdalny wolny	(Vullo et al., 2006)
PDISORDER[1,3]	brak uporządkowania	lokalny	Softberry, Inc.
POODLE-L[1,3]	brak uporządkowania	zdalny szybki	(Hirose et al., 2007)
POODLE-S[1,3]	brak uporządkowania	zdalny wolny	(Shimizu et al., 2007)
PrDOS[1,3]	brak uporządkowania	zdalny wolny	(Ishida i Kinoshita, 2007)
iPDA[1,3]	brak uporządkowania	zdalny wolny	(Su et al., 2007)
Metadisorder	brak uporządkowania	lokalny	(Schlessinger et al., 2009)
SPINE-D	brak uporządkowania	lokalny	(Zhang et al., 2012)
GeneSilicoMetaDisorder	**brak uporządkowania**	**lokalny**	**(Kozlowski i Bujnicki, 2012)**
NetSurfP	RSA	lokalny	(Petersen et al., 2009)
SOPRANO	RSA	lokalny	(Kirschner i Frishman, 2008)
ACCpro	RSA	lokalny	(Pollastri et al., 2002)
JNET	RSA	lokalny	(Cole et al., 2008)
SPINE	RSA	lokalny	(Dor i Zhou, 2007)
SPINE X	RSA	lokalny	(Faraggi et al., 2012)
SABLE	RSA	lokalny	(Wagner et al., 2005)
WESA	RSA	lokalny	(Tjong et al., 2007)
Phobius	helisy transbłonowe	lokalny	(Kall et al., 2007)
MEMSAT3	*helisy transbłonowe*	*lokalny*	*(Jones, 2007)*
TMPred	*helisy transbłonowe*	*lokalny*	*(Hofmann i Stoffel, 1993)*
HMMTOP	helisy transbłonowe	lokalny	(Tusnady i Simon, 2001)
DAS	helisy transbłonowe	lokalny	(Cserzo et al., 2004)
MINNOU	helisy transbłonowe	lokalny	(Cao et al., 2006)
TMHMM2.0	*helisy transbłonowe*	*lokalny*	*(Krogh et al., 2001)*
PRODIV	helisy transbłonowe	zdalny wolny	(Viklund i Elofsson, 2004)
PRO	helisy transbłonowe	zdalny wolny	(Viklund i Elofsson, 2004)
S_TMHMM	helisy transbłonowe	zdalny wolny	(Viklund i Elofsson, 2004)
SCAMPI-seq	helisy transbłonowe	zdalny wolny	(Bernsel et al., 2008)
SCAMPI-msa	helisy transbłonowe	zdalny wolny	(Bernsel et al., 2008)
OCTOPUS	helisy transbłonowe	zdalny wolny	(Viklund i Elofsson, 2008)
TOPCONS	helisy transbłonowe	zdalny wolny	(Bernsel et al., 2009)
Proteus	helisy transbłonowe	lokalny	(Montgomerie et al., 2008)
COILS	splecione helisy	lokalny	(Lupas et al., 1991)
PAIRCOIL	splecione helisy	lokalny	(Berger et al., 1995)
MARCOIL	splecione helisy	lokalny	(Wolf et al., 1997)
PCOILS	splecione helisy	lokalny	(Gruber et al., 2005)
CAST	splecione helisy	lokalny	(Promponas et al., 2000)
NCOILS	splecione helisy	lokalny	-
DP-BIND	DNA-białko	zdalny wolny	(Hwang et al., 2007)
DISIS	DNA-białko	lokalny	(Ofran et al., 2007)
BindN	RNA-białko	zdalny szybki	(Wang i Brown, 2006)
BindN	DNA-białko	zdalny szybki	(Wang i Brown, 2006)
BindN-RF	DNA-białko	zdalny wolny	(Wang et al., 2009)

Nazwa programu	Przewidywana cecha	Kategoria	Referencja
BindN+	DNA-białko	zdalny wolny	(Wang et al., 2010)
BindN+	RNA-białko	zdalny wolny	(Wang et al., 2010)
DBindR	DNA-białko	zdalny szybki	(Wu et al., 2009)
NAPS	RNA-białko	zdalny szybki	(Carson et al., 2010)
NAPS	DNA-białko	zdalny szybki	(Carson et al., 2010)
PPRInt	RNA-białko	zdalny wolny	(Kumar et al., 2007)
PiRaNhA	RNA-białko	zdalny szybki	(Murakami et al., 2010)
RNABindR	RNA-białko	zdalny szybki	(Terribilini et al., 2007)
cons3best	**RNA-białko**	**lokalny**	**(Puton et al., 2012)**
CYS_REC	mostki dwusiarczkowe	zdalny szybki	Softberry, Inc.
DiANNA	mostki dwusiarczkowe	zdalny szybki	(Ferre i Clote, 2006)
DIpro2	mostki dwusiarczkowe	lokalny	(Cheng et al., 2006a)
DISULFIND	mostki dwusiarczkowe	lokalny	(Ceroni et al., 2006)
DBCP	mostki dwusiarczkowe	zdalny wolny	(Lin i Tseng, 2010)
PSI-BLAST[2,3]	*rozpoznawanie zwoju*	*lokalny*	*(Altschul et al., 1997)*
GenTHREADER[2,3]	*rozpoznawanie zwoju*	*lokalny*	*(Jones, 1999)*
COMPASS	rozpoznawanie zwoju	lokalny	(Sadreyev et al., 2007)
PRC	rozpoznawanie zwoju	lokalny	(Madera, 2008)
COMA	rozpoznawanie zwoju	lokalny	(Margelevicius i Venclovas, 2010)
CS-BLAST	rozpoznawanie zwoju	lokalny	(Biegert i Soding, 2009)
HHsearch[2,3]	rozpoznawanie zwoju	lokalny	(Soding, 2005)
HHblits[4]	rozpoznawanie zwoju	lokalny	(Remmert et al., 2012)
FFAS[2,3]	*rozpoznawanie zwoju*	*lokalny*	*(Rychlewski et al., 2000)*
Sp3	rozpoznawanie zwoju	lokalny	(Zhou i Zhou, 2005)
Phyre[2,3]	rozpoznawanie zwoju	zdalny wolny	(Bennett-Lovsey et al., 2007)
PCONS[2,3]	*rozpoznawanie zwoju*	*lokalny*	*(Wallner i Elofsson, 2005)*
JMBRANK	rozpoznawanie zwoju	lokalny	-
Consens3d	rozpoznawanie zwoju	lokalny	-
fastSCOP	inne	zdalny szybki	(Tung i Yang, 2007)
EnzyPred	inne	zdalny szybki	(Shen i Chou, 2007)
MODELLER	inne	lokalny	(Sali et al., 1995)
Swiss PDB Viewer	inne	lokalny	(Guex i Peitsch, 1997)
HMMER	*inne*	*lokalny*	*(Finn et al., 2011)*
DSSP	inne	lokalny	(Kabsch i Sander, 1983)
SCWRL	*inne*	*lokalny*	*(Dunbrack, 1999)*
FILTREST3D	inne	zdalny wolny	(Gajda et al., 2010)
Frankenstein3D	inne	zdalny wolny	(Kosinski et al., 2005)
MetaMQAP	inne	lokalny	(Pawlowski et al., 2008)
AmIgoMR	inne	zdalny szybki	(Pawlowski i Bujnicki, 2012)
MetaMQAPclust	inne	lokalny	(Pawlowski i Bujnicki, 2012)

6.1.1. Struktura serwisu

Metaserwer jest serwisem internetowym, w obrębie którego można wyróżnić trzy warstwy (ryc. 10). Najbardziej widoczną warstwą jest interfejs użytkownika, czyli ta część serwisu, do której mają dostęp użytkownicy. Została ona napisana w serwerze aplikacji Zope, który pozwala w efektywny sposób zarządzać serwisem oraz generować strony HTML w sposób dynamiczny. Najgłębszą warstwę serwisu stanowią napisane w języku Python skrypty pozwalające na uruchomienie programów bioinformatycznych składających się na metaserwer. Należą do niej tu także skrypty zarządzające kolejkowaniem zapytań oraz skrypty których zadaniem jest ujednolicenie pierwotnego formatu wyników do formy, jaką przyjęto dla danej kategorii programów. Ostatnią warstwą, spajającą dwie powyższe jest warstwa baz danych MySQL. W niej przechowywane są wszelkie dane dotyczące zarówno użytkowników jak i samych zapytań. W bazach przechowywane są także wszystkie pierwotne wyniki programów, wliczając w to przyrównania i modele homologiczne.

Ze względu na specyfikę serwisu i liczbę programów włączonych do metaserwera do wydajnego działania serwisu niezbędne było ulokowanie poszczególnych jego elementów na dwóch oddzielnych komputerach. Bazy danych oraz interfejs użytkownika zlokalizowane są na względnie słabej maszynie (4 CPU, 16 GB RAM), natomiast większość zadań złożonych obliczeniowo dedykowane jest na mocniejszy komputer (48 CPU, 96 GB RAM). Rozwiązanie takie zapewnia niezawodność działania całości bez względu na obciążenie i liczbę zadań wysłanych przez użytkowników.



Ryc. 10. Schemat budowy metaserwera GeneSilico do rozpoznawania zwoju.

6.1.2. Interfejs użytkownika

Jak wspomniano wyżej panel użytkownika został napisany w serwerze aplikacji Zope. Pozwala on w efektywny sposób zbudować system zarządzania treścią i przydzielić odpowiednie prawa poszczególnym użytkownikom. Administrator systemu ma dostęp do panelu administratora (ryc. 11), który służy do zarządzania całym serwisem. Z tego poziomu można tworzyć podgrupy i ograniczać lub poszerzać prawa użytkowników (np. umożliwiając oglądanie i modyfikowanie wyników zapytań danej podgrupy). W chwili obecnej istnieją trzy typy użytkowników: *guest* – użytkownik nienależący do żadnej grupy (domyślne ustawienie) mający dostęp jedynie do własnych zapytań, grupa *lab* skupiająca użytkowników z Laboratorium Bioinforamatyki i Bioinżynierii Białka, mających wgląd w zapytania innych członków grupy oraz *superusers*, czyli użytkownicy mający wgląd we wszystkie dane, włączając w to panel administratora. W miarę potrzeby istnieje możliwość wydzielenia innych podgrup.

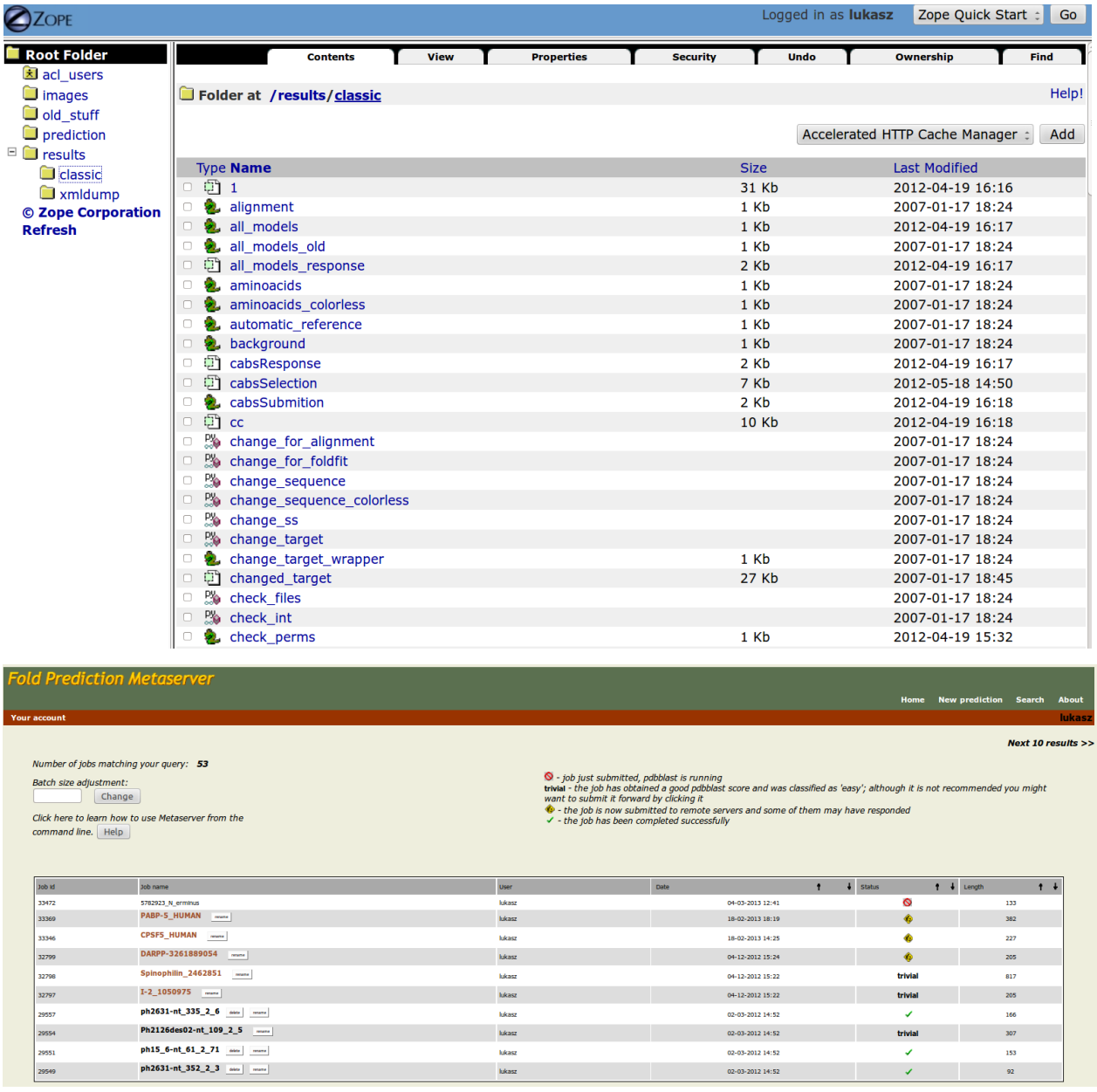
Ze względów bezpieczeństwa serwis jest dostępny jedynie po zarejestrowaniu, ale rejestracja jest bezpłatna, szybka i w pełni zautomatyzowana (użytkownik otrzymuje swój login i hasło na wskazany adres email). Po rejestracji użytkownik otrzymuje dostęp do panelu, który umożliwia przeglądanie historii swoich zapytań oraz posiada funkcję przeszukiwania poprzednich wyników (możliwe jest stosowanie wyrażeń regularnych np. zapytanie „*p53*” w polu „*Job name”* zwróci wszystkie zapytania, których tytuł zawiera słowo „p53”, zakładając, że użytkownik jest ich właścicielem lub ma prawo do ich odczytu) – ryc. 11. Dodatkowo każdy użytkownik ma dostęp do wyników zapytań pewnych szczególnych użytkowników takich jak „*student*” czy „*casp*” (ten ostatni przechowuje wyniki dla wszystkich zapytań związanych z sekwencjami, które były celem badawczym eksperymentu CASP, oznacza to możliwość przejrzenia wyników poczynając od edycji CASP6 (rok 2004) do CASP10 (rok 2012); dostęp do tego typu danych jest niezwykle użyteczny, ponieważ na ich podstawie możliwe jest konstruowanie różnego typu zbiorów testowych, które można wykorzystać w czasie tworzenia nowych metod bioinformatycznych).

Przeszukiwanie historii zapytań jest szczególnie ważne, ponieważ wyniki poszczególnych programów uruchamianych przez metaserwer dodawane są w miarę ich dostępności, a pojawienie się wyników ze wszystkich metod może trwać w skrajnych przypadkach nawet kilka dni. Mniej więcej połowa programów jest w stanie zwrócić wynik w relatywnie krótkim czasie (w ciągu kilku minut). Pozostałe programy działają dosyć wolno, co

najczęściej wynika z dwóch przyczyn. Niektóre programy lokalne zostały napisane w sposób uniemożliwiający zrównoleglenie obliczeń (np. COMA, GenThreader) opierając się jednocześnie na przeszukiwaniu dużych baz danych. W ten sposób pojedyncze uruchomienie programu może wymagać 1–2 godzin. Drugi problem jest specyficzny dla serwisów internetowych. Nawet jeśli serwis internetowy działa w miarę szybko, istnieje ryzyko, że komunikacja między serwisem zewnętrznym a metaserwerem zawiedzie i niezbędne będzie ponowne wysłanie zapytania (powody takiego zdarzenia mogą być różne np. zerwanie połączenia z Internetem po którejkolwiek stronie, nieodesłanie przez serwer wyniku na wskazany email, błąd w przetwarzaniu wyniku po zmianach na stronie internetowej serwisu zewnętrznego itd.).

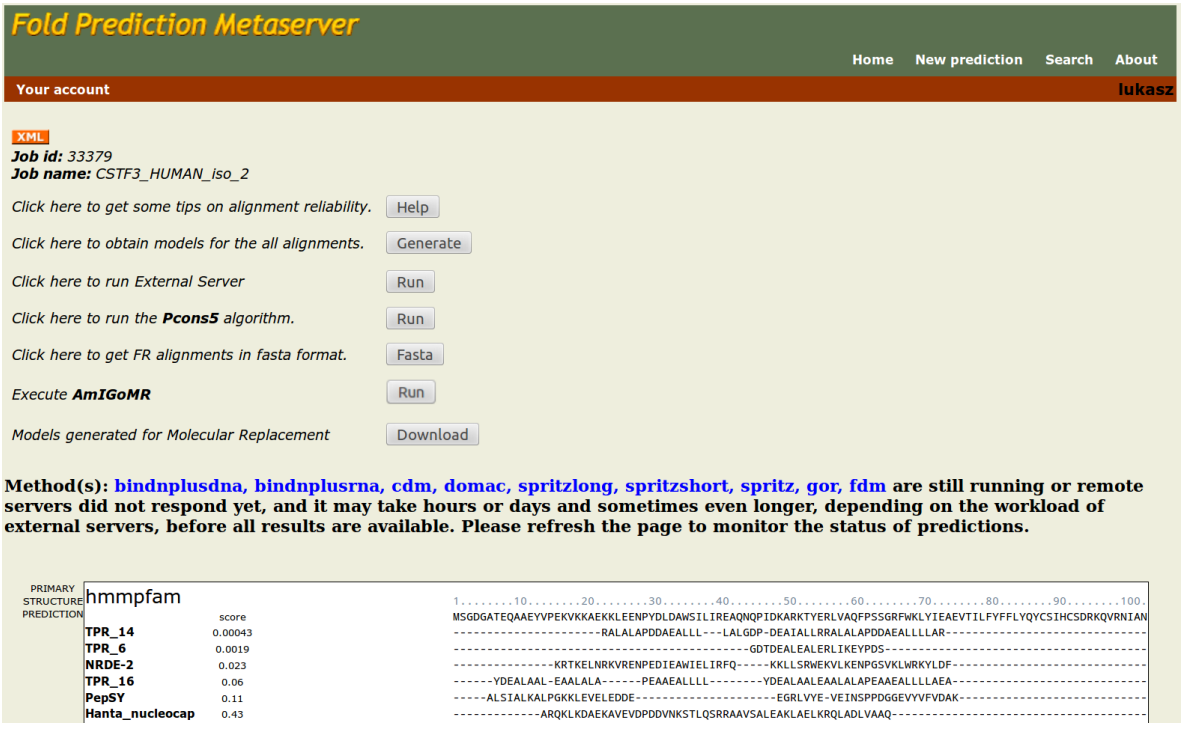
Dla ułatwienia pracy z metaserwerem użytkownik jest powiadamiany przez email o zmianach w statusie wysłanego zapytania. Najpierw, po wykonaniu obliczeń przez programy lokalne, do użytkownika wysyłany jest email z informacją, że wyniki metod lokalnie zainstalowanych są już dostępne. W tym momencie dochodzi do automatycznej klasyfikacji poziomu trudności zapytania. Jeśli program PSI-BLAST znajduje szablony z wartością *e-value* poniżej określonego progu to zapytanie klasyfikowane jest jako łatwe czyli przyjmuje status „*trivial*" – oznacza to, że programy zewnętrzne nie będą uruchamiane dla tego zapytania chyba, że użytkownik zażyczy sobie inaczej (poprzez kliknięcie na napis „*trivial*" status zmieniany jest na status o nazwie „*some*" (oznaczony żółtą ikoną na ryc. 6.1.2.), co powoduje wysłanie sekwencji użytkownika do serwisów zewnętrznych). Jeśli program PSI-BLAST nie znajduje dobrych szablonów status „*trivial*" automatycznie zmieniany jest na status „*some*". W momencie gdy metaserwer otrzymuje wszystkie wyniki łącznie z wynikami z serwerów zewnętrznych status przyjmuje wartość „*all*" (zielona ikona na ryc. 11). Wtedy też wysyłany jest do użytkownika kolejny email informujący o zakończeniu analiz. Ze względu na dużą liczbę programów, które są włączone do metaserwera istnieje wysokie ryzyko, że przynajmniej jeden lub kilka serwisów chwilowo nie będzie działać. Wchodząc na podstrony wyników użytkownik może łatwo sprawdzić, czy w danym momencie brakuje jeszcze wyników jakichś metod i ocenić, czy brak określonych przewidywań w znaczący sposób może wpłynąć na interpretacje wyników. Przykładowo, jeśli widzimy, że ciągle brakuje wyników dla serwisu PrDOS, który przewiduje brak uporządkowania, ale wyniki pozostałych metod są już dostępne, to możemy uznać, że brak wyników jednej metody nie wypłynie znacząco na wynik konsensusowy,

ponieważ budowany jest on na podstawie wyników z kilku lub kilkunastu innych programów, więc jego jakość nie powinna być znacząco obniżona.



Ryc. 11. Panel użytkownika metaserwera GeneSilico. U góry przedstawiono panel administratora, który pozwala na dogodny dostęp i modyfikację treści poprzez skrypty w języku Python oraz dynamiczne szablony stron HTML wykorzystujące znaczniki DTML (ang. *Document Template Markup Language*) i TAL (ang. *Template Attribute Language*); na dole przedstawiono wygląd panelu użytkownika, który pozwala na pełny wgląd w historię zapytań, pokazuje ich status, datę rozpoczęcia i inne użyteczne informacje.

Cechą szczególną metaserwera jest sposób prezentacji wyników, którego zadaniem jest ułatwienie interpretacji wyników analizy. Wynik programów każdej grupy przedstawione są w jednolitym formacie tekstowym, dzięki czemu można łatwo je porównać ze sobą nawzajem. Poza tym dla każdej grupy metod zgłaszających ten sap typ wyników, wyliczany jest konsensus, który pozwala na ocenę zbieżności przewidywań (zazwyczaj przewidywania, które nawzajem się ze sobą zgadzają, są także najbardziej wiarygodne). Do każdej grupy programów dodany jest link oznaczony symbolem „?”. Kieruje on do strony, na której wyjaśnione jest jak należy interpretować wyniki oraz podane są referencje do publikacji i stron internetowych metod składowych). Na samej górze strony z wynikami (ryc. 12) użytkownik może znaleźć podstawowe informacje (nazwę zapytania, jego numer) oraz informację dotyczącą aktualnego stanu zapytania (na niebiesko podana jest lista metod, które na dany moment jeszcze nie zwróciły wyników). Ponadto użytkownik ma w tej sekcji dostęp do kilku przycisków, które pozwalają na zbiorcze pobieranie danych np. wszystkich modeli, lub wszystkich przyrównań w formacie fasta). Część przycisków łączy metaserwer z programami zewnętrznymi (po ich naciśnięciu odpowiednie dane wysyłane są automatycznie do serwisów zewnętrznych, np. do wymienionego wcześniej programu AmIGoMR).



Ryc. 12. Górna część przykładowej strony z wynikiem na metaserwerze GeneSilico. Prezentowane są ogólne informacje, status zapytania oraz przyciski pozwalające włączać określone programy zewnętrzne. Ponadto możliwe jest pobranie modeli i przyrównań.

6.1.2.1. Programy do rozpoznawania zwoju

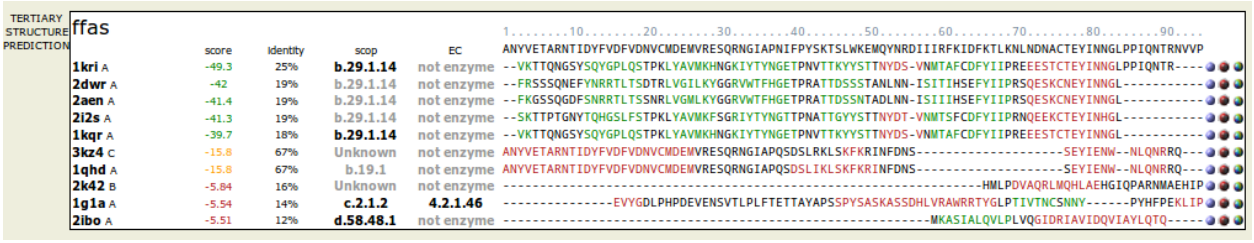
Programy do rozpoznawania zwoju mają za zadanie znaleźć najlepszy możliwy szablon do modelowania homologicznego badanego białka. Domyślnie przeszukują one bazy struktur zbudowane na podstawie bazy PDB. W skład metaserwera wchodzi aktualnie 12 różnych programów do rozpoznawania zwoju. Wszystkie z wyjątkiem programu Phyre, zainstalowane są lokalnie. Najprostsze z nich to metody typu sekwencja-profil (np. PSI-BLAST, GenTHREADER, COMPASS), w przypadku których sekwencja porównywana jest z bazą profili PSSM. Dodatkowo zamiast pojedynczej sekwencji może być użyte przyrównanie MSA konstruowane we wstępnym etapie analizy poprzez przeszukanie bazy sekwencji (np. nr lub UniProt). Dopiero ono jest używane w trakcie głównego przeszukiwania bazy profili. Różnice pomiędzy poszczególnymi metodami dotyczą algorytmów konstruowania przyrównania (np. w przypadku PSI-BLAST wykorzystywany jest algorytm Smitha–Watermana (Smith i Waterman, 1981)) oraz rodzaju użytych baz danych (część programów np. COMA posiada specyficzne bazy danych o dosyć nietypowym formacie, które są częściej lub rzadziej aktualizowane przez autorów programów).

Następną grupę stanowią programy typu profil-profil, w których z sekwencji-zapytania tworzony jest profil lub model HMM i następnie za jego pomocą przeszukiwana jest baza profili lub modeli HMM (przykładowe programy to FFAS oraz HHsearch). Każdy z programów stosuje własny system oceny jakości przyrównania zapytania i potencjalnych szablonów. Aby ułatwić interpretacje wyników, wyświetlane liczby opisujące wartość oceny kolorowane są według następującego klucza: kolor zielony odpowiada przyrównaniom dobrym, kolor pomarańczowy oznacza średnią jakość, natomiast na czerwono zaznaczone są wartości dla przyrównań słabej jakości. Szczegółowe wartości progów odcięcia jakości dla poszczególnych metod, ich klasyfikacja i użyte bazy podane zostały w tabeli 3. Domyślnie prezentowane jest do 10 najlepszych szablonów w następującej postaci: kod PDB szablonu, jakość przyrównania, procent identyczności między sekwencją a szablonem, klasyfikacja SCOP, numer enzymu EC, przyrównanie pokolorowane według struktury drugorzędowej, przyrównania sekwencja-szablon w formatach PIR i FIT oraz modele wygenerowane procedurą „*crude*” lub za pomocą programu MODELLER (ryc. 13). Domyślnie generowane są modele typu „*crude*”, ponieważ ich budowa jest szybsza. Wygenerowanie modelu programem MODELLER, choć zwykle daje lepsze

wyniki, jest bardziej kosztowne obliczeniowo (pojedyncze zapytanie użytkownika daje około 100 przyrównań sekwencja-szablon, dla których metaserwer automatycznie buduje modele), więc ten typ generowania modeli został zarezerwowany dla szczególnych zadań np. dla użytkownika „*casp*”.

Tabela 3. Charakterystyka programów do wykrywania zwoju białka. W przypadku progów odcięcia jakości kolorem zielonym zaznaczono próg powyżej (lub poniżej) którego szablon uznaje się za wiarygodny. Kolorem pomarańczowym oznaczono przedział odzwierciedlający średnią jakość przyrównania, zaś kolor czerwony oznacza próg poniżej (lub powyżej) którego szablon uznaje się za mało wiarygodny. Progi odcięcia przyjęto zgodnie z zaleceniami autorów programów lub ustalono empirycznie w przypadku braku takich zaleceń.

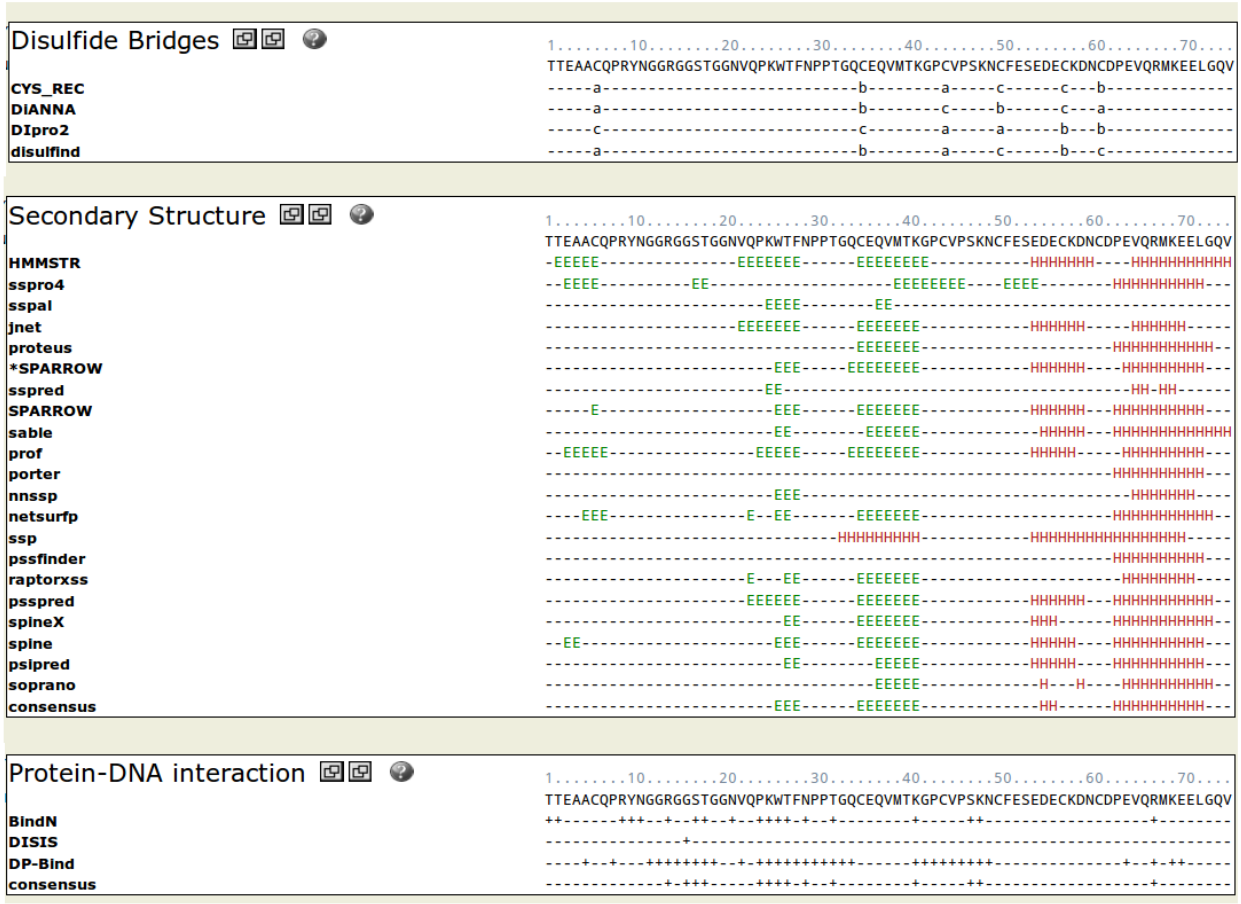
Nazwa programu	**Przeszukiwana baza danych**	**Typ programu**	**Progi odcięcia jakości szablonów**
PSI-BLAST a) pdbblast b) blastp	nr90, CULLPDB z opcją -f bez opcji -f	sekwencja-profil (PSSM)	<=2e-06 <= 0.023 > 0.023
GenTHREADER	uniref90, CATH i inne specyficzne bazy generowane przez autorów	sekwencja-profil (PSSM)	<= 0.001 <= 0.01 > 0.01
COMPASS	nr90, pdb70	sekwencja-profil (PSSM)	<=2e-06 <= 0.023 > 0.023
PRC	baza generowana przez autorów programem HMMER	profil-profil (HMM)	<=2e-06 <= 0.023 > 0.023
COMA	nr90, profile pdb40 (dostarczane przez autorów)	profil-profil (własny format)	<=2e-06 <= 0.023 > 0.023
CS-BLAST	CULLPDB	sekwencja-profil (PSSM)	<=2e-06 <= 0.023 > 0.023
HHblits	nr20, pdb70	sekwencja-profil (HMM)	>= 95.0 >= 80.00 < 80.00
HHsearch	nr20 i pdb70 lub nr20 i CDD	profil-profil (HMM)	>= 95.0 >= 80.00 < 80.00
FFAS	specyficzna baza dostarczana przez autorów zbudowana na podstawie PDB	profil-profil (PSSM)	<= -34.5 <= -8.5 > 8.50
Sp3	nr90, SCOP i bazy pochodne PDB dostarczane przez autorów	profil-profil	<= -150 <= -80 > -80.00
Phyre	nr, swissprot	profil-profil (HMM)	<= 0.085 <= 0.27 > 0.27
HMMER	PFAM	profil-profil (HMM)	<=2e-06 <= 0.023 > 0.023
PCONS	brak, program ocenia jakość modeli wygenerowanych na podstawie szablonów wykrytych przez poprzednie metody	-	>= 2.17 >= 1.030 < 1.030



Ryc. 13. Przykład wyniku z programu do rozpoznawania zwoju (program FFAS). Standardowo prezentowane jest do 10 najlepszych szablonów. W pierwszej kolumnie znajdują się kody PDB szablonów, dalej informacje o jakości przyrównań, procent identyczności sekwencji-zapytania i szablonu (wyliczany jedynie dla fragmentów które sobie odpowiadają), klasyfikacja SCOP danego fragmentu szablonu i informacja o ewentualnej klasyfikacji szablonu w bazie enzymów EC (kolor czarny oznacza, że dane pochodzą z pliku PDB szablonu, natomiast kolor szary oznacza, że dane przewidziano za pomocą programów fastSCOP i EnzyPred). Następnie prezentowane jest samo przyrównanie szablonu do sekwencji docelowej (dla ułatwienia oceny jakości przyrównania sekwencja szablonu jest pokolorowana zgodnie ze strukturą drugorzędową pochodzącą z szablonu). Po prawej stronie znajduje się panel z trzema przyciskami (dla każdego przyrównania) w formie „kulek". Pozwalają one pobrać przyrównanie w formacie PIR (domyślny format programu MODELLER) – „kulka" niebieska, przyrównanie w formacie FIT (format programu Swiss PDB Viewer) – „kulka" czerwona i model wygenerowany przez procedurę „*crude*" („kulka" zielona).

6.1.2.2. Programy do analizy innych własności białek

Oprócz programów do wykrywania zwoju białka metaserwer pozwala uruchomić cały szereg programów przewidujących takie właściwości białek jak struktura drugorzędowa, obecność helis transbłonowych, sekwencji sygnalnych, mostków dwusiarczkowych, struktur splecionych helis, dostępność reszt aminokwasowych dla rozpuszczalnika, domeny oraz wewnętrzne nieuporządkowanie (patrz tabela 2). Jedną z ważniejszych cech wyróżniającą metaserwer GeneSilico na tle innych tego typu serwisów jest sposób prezentacji wyników. Choć programy z każdej kategorii przewidują występowanie tej samej cechy białka, praktycznie każdy program zwraca wyniki w innym formacie. W takiej sytuacji porównanie wyników kilku programów jest trudne, a czasem wręcz niemożliwe bez umiejętności programowania. Z tego powodu ważne jest przekształcenie wszystkich wyników do jednorodnego formatu, nawet jeśli oznacza to częściową utratę informacji. Przykładowo wyniki programów przewidujących strukturę drugorzędową w metaserwerze zredukowane są trzech stanów (H – helisa α, E – wstęga β i „-" – inne), nawet jeśli program jest w stanie przewidzieć np. osiem typów struktur drugorzędowych według klasyfikacji DSSP (ryc. 14). Ta częściowa utrata informacji rekompensowana jest możliwością bezpośredniego porównania wyników różnych metod i stworzenia konsensusu.



Ryc. 14. Wyniki programów przewidujących mostki dwusiarczkowe, strukturę drugorzędową i oddziaływanie białka z DNA na przykładzie białka należącego do rodziny Kunitza pochodzącego z kleszcza pospolitego (*Ixodes ricinus*). W przypadku mostków dwusiarczkowych resztom cystein tworzącym potencjalny mostek przypisane są takie same litery (np. program CYS_REC przewidział mostek siarczkowy między resztami o numerze 6 i 44, 35 i 61 oraz 50 i 57). Zestawienie wyników wielu programów w jednorodnym formacie pozwala łatwo wyeliminować niepewne wyniki. Przykładowo większość programów do przewidywania struktury drugorzędowej (środkowy panel) przewiduje w regionie 35-41 występowanie wstęgi β, jednak niektóre programy, prawdopodobnie, mylnie przypisują do tego regionu występowanie helisy α (program ssp) lub całkowity brak struktury drugorzędowej (pssfinder, porter i nnssp).

6.1.3. Struktura baz danych

Ze względów na wydajność i bezpieczeństwo dane użytkowników znajdują się w oddzielnej bazie danych zawierającej w sumie 11. tabel. W czasie rejestracji użytkownik proszony jest o podanie danych takich jak imię i nazwisko, instytucja naukowa, kraj pochodzenia itp. Dane te pozwalają na kontakt w celach informacyjnych (powiadamianie o dodaniu nowych metod, informacje o pracach modernizacyjnych itd.) lub w sytuacjach awaryjnych. W styczniu 2013 roku liczba użytkowników metaserwera GeneSilico przekroczyła 2300 (w 2008 roku w

momencie w którym autor rozprawy przejął obowiązki związane z rozwojem serwisu użytkowników tych było niecałe 500.).

Dane dotyczące zapytań (pierwotne wyniki poszczególnych metod, modele homologiczne, przyrównania) przechowywane są w innej bazie danych. Każda z kategorii metod reprezentowana jest w formie oddzielnej tabeli. Jedynym wyjątkiem są tu metody do rozpoznawania zwoju, z których każda posiada oddzielną tabelę (rozwiązanie takie jest częściowo sprzeczne z zasadą normalizacji baz danych, ale ułatwia tworzenie kopii zapasowych; tabele metod do wykrywania zawierają najwięcej danych, ponieważ w nich przechowywane są przyrównania i modele). Ponadto w bazie tej istnieje szereg tabel pomocniczych np. tabele przechowujące numery EC czy kody PDB wraz z ich objaśnieniami. W sumie baza danych zbudowana jest z 51. tabel. W styczniu 2013 baza ta miała wielkość 34 GB, z czego ponad 90% danych stanowiły modele.

Szczególnie ważną tabelą jest tabela kolejki. Pozwala ona wydajnie zarządzać zapytaniami do poszczególnych programów i serwisów zewnętrznych. Domyślnie zapytania kolejkowane są według kolejności przysłania przez użytkowników (wyjątkiem jest użytkownik „*casp*”, którego zapytania mają najwyższy priorytet i wysyłane są jako pierwsze). Każda z metod ma swoją własną kolejkę, z której sekwencje pobierane są jedna po drugiej. Kolejne zapytania wysyłane są dopiero po otrzymaniu wyniku. Dodatkowo w przypadku braku odpowiedzi po określonym czasie (zwykle 24 godziny) zapytanie jest uruchamiane ponownie. Kolejki poszczególnych metod działają niezależnie od siebie, dzięki czemu brak odpowiedzi z jednego programu nie blokuje wysyłania zapytań do innych programów.

6.1.4. Skrypty zarządzające uruchamianiem programów składowych

Najgłębszą warstwę metaserwera stanowią skrypty w języku Python odpowiadające za przetwarzanie zapytań użytkowników, uruchamianie poszczególnych programów oraz przetwarzanie ich wyników. Ze względu na dużą liczbę programów i ich różnorodność każdy z programów zarządzany jest przez oddzielny moduł. Pozwala to zachować niezależność poszczególnych komponentów metaserwera. Czasem, sytuację komplikuje fakt, że pliki wynikowe pochodzące z jednych programów są niezbędne do uruchomienia innych. Duża część programów (np. program do przewidywania struktury drugorzędowej Jnet) wymaga pliku wynikowego z programu PSI-BLAST. W takich przypadkach niezbędne jest uruchomienie

pewnych programów w określonej kolejności. W standardowej sytuacji programy są jednak niezależne i w teorii mogą być uruchomione w oddzielnych wątkach niezależnie od siebie. Niestety, ze względu na duże wymagania obliczeniowe część programów uruchamiane jest w praktyce sekwencyjnie, mimo wykorzystania względnie mocnego komputera (48 CPU, 96 RAM).

Od strony technicznej każdy moduł zbudowany jest z 4 części: pierwsza część odpowiada za sprawdzenie stanu kolejki. W przypadku niezerowego stanu kolejki wykonywana jest druga część modułu, czyli właściwe uruchomienie programu składowego (na tym etapie następuje zablokowanie możliwości uruchamiania nowych instancji modułu do momentu zakończenia działania programu). Trzecia część modułu odpowiada za przetworzenie wyniku do przyjętego formatu. Ostatnia część modułu aktualizuje bazy danych o uzyskany wynik oraz wykonuje czynności końcowe (usunięcie blokady, usunięcie plików tymczasowych itp.).

Opisana wyżej struktura modułów dotyczy jedynie programów lokalnie zainstalowanych i serwisów zewnętrznych zwracających wynik bezpośrednio na stronie internetowej. Oddzielna kategorię programów stanowią serwisy internetowe zwracające wynik na wskazany adres email. W tym przypadku niezbędne jest zastosowanie obsługi zapytania o bardziej złożonej strukturze. Sprawdzanie stanu kolejki i wysyłanie sekwencji na serwisy zewnętrzne zarządzane jest przez jeden skrypt. Natomiast za obróbkę wyników i ich wstawianie do bazy danych odpowiada inny skrypt, którego głównym zadaniem jest wydobycie odpowiedzi serwisu ze skrzynki email. Rozdzielenie tych zadań jest konieczne, ponieważ uruchomienie programu i otrzymanie wyniku jest rozdzielone w czasie, którego długości nie da się przewidzieć.

Wszystkie skrypty nadrzędne uruchamiające poszczególne moduły oraz przetwarzające zawartość skrzynki email oraz skrypty aktualizujące biologiczne bazy danych uruchamiane są przez zewnętrzny program *crontab*, który pozwala uruchamiać dowolny program uruchamiany w środowisku systemu Unix w ustalonych odstępach czasu.

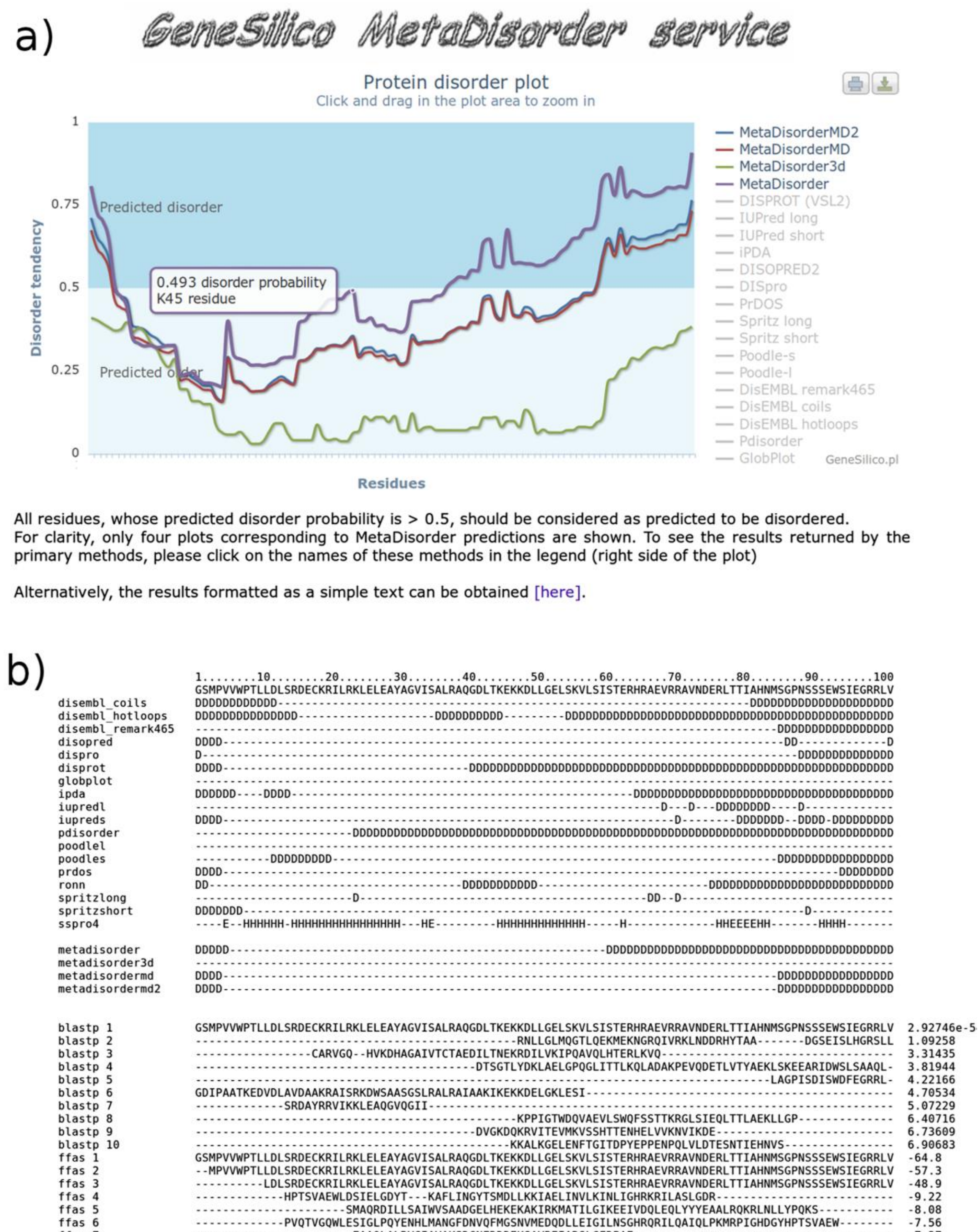
6.2. Przewidywanie wewnętrznego nieuporządkowania struktury

W ramach realizacji niniejszego projektu skonstruowano serię metod służących do przewidywania wewnętrznego nieuporządkowania struktury białka. Celem było otrzymanie takiej metody, która uzyskałaby możliwie najlepsze wyniki w porównaniu do innych istniejących metod, w szczególności pod względem oceny w ramach „zawodów CASP". Cel badawczy obejmował przetestowanie użyteczności różnych technik uczenia maszynowego do realizacji tego zadani oraz zbadanie wpływu użycia różnych danych wejściowych (np. na ile dołączenie wyniku z przewidywania struktury drugorzędowej może poprawić wynik przewidywania wewnętrznego nieuporządkowania). Podejście takie umożliwiło stworzenie kilku wariantów programu, które wraz ze stopniem skomplikowania zyskiwały coraz lepszą skuteczność w przewidywaniu wewnętrznego nieuporządkowania struktury białek, względem różnych zestawów danych testowych.

6.2.1. MetaDisorder – meta-metoda oparta na innych programach do przewidywania regionów wewnętrznie nieuporządkowanych

MetaDisorder to program opierający się na klasycznej koncepcji meta-metody. Program wykorzystuje kilkanaście innych programów przewidujących występowanie regionów wewnętrznie nieuporządkowanych: DisEMBL, GLOBPLOT, DISOPRED2, DISPROT (VSL2B), IUPred (uruchamiany w dwóch trybach dostosowanych do przewidywania długich i krótkich regionów wewnętrznie nieuporządkowanych), DISpro, RONN, SPRITZ, PDISORDER, POODLE-L, POODLE-S, PrDOS oraz iPDA (więcej informacji na temat tych programów można znaleźć w „Materiałach i metodach" w rozdziale 5.3.3.3 oraz w tabeli 2).

Program MetaDisorder jest dostępny jako jedna z metod uruchamianych przez metaserwer GeneSilico oraz jako niezależny serwis internetowy pod adresem http://genesilico.pl/metadisorder/. Podobnie jak w przypadku metaserwera GeneSilico, wprowadzono ograniczenia dotyczące danych wejściowych (limit wysłanych zapytań do dziesięciu zapytań dla jednego użytkownika na dzień, długość sekwencji od 40 do 990 aminokwasów). W przypadku serwisu internetowego wynik prezentowany jest w postaci graficznej, przyjaznej dla użytkownika oraz w postaci tekstowej ułatwiającej komputerową obróbkę wyników (ryc. 15).



Ryc. 15. Interfejs serwisu internetowego MetaDisorder: a) graficzna postać wyniku przyjazna dla użytkownika (wykres można dowolnie skalować, zapisywać w formie obrazu, itp.), ponadto przewidywania wszystkich metod dostępne są do pobrania w jednorodnym formacie CASP; b) prosty format tekstowy.

6.2.1.1. Definicja wewnętrznego nieuporządkowania

Jak wspomniano wcześniej, nie ma jednej powszechnie akceptowanej definicji regionów wewnętrznie nieuporządkowanych. Istnieją różne definicje m.in. uzależnione od rodzaju użytej metody doświadczalnej, które zgadzają się ze sobą jedynie częściowo (np. regiony wyznaczone

jako nieuporządkowane w obrębie określonego białka za pomocą techniki NMR nie zawsze zgadzają się z regionami wyznaczonymi za pomocą innej techniki). Ze względów praktycznych na potrzeby realizacji niniejszego projektu badawczego przyjęto definicję obowiązującą w ramach eksperymentu CASP, gdzie za regiony wewnętrznie nieuporządkowane uznaje się te, które oznaczone są w plikach PDB znacznikiem REMARK465 (reszty aminokwasowe, które są obecne w sekwencji białka, ale nie ma dla nich koordynatów), którą rozszerzono o regiony, oznaczone jako nieuporządkowane w bazie danych DisProt. W ten sposób definicja braku uporządkowania uwzględnia dane doświadczalne otrzymane wieloma różnymi technikami, takimi jak krystalografia rentgenowska, NMR, elektroforeza SDS-PAGE, SAXS, spektrometria mas i inne. Podejście takie pozwala także uniknąć wprowadzenia błędu systematycznego polegającego na ograniczeniu badań do białek o znanej strukturze trzeciorzędowej, w sytuacji gdy zadaniem metody jest m.in. wykrywanie nieuporządkowania w białkach, które w dużej części lub w całości pozbawione są struktury, czyli ich struktura właściwie nie ma szans na znalezienie się w bazie danych PDB.

6.2.1.2. Zbiór testowy

Mając na względzie przyjętą definicję regionów wewnętrznie nieuporządkowanych, skonstruowano odpowiedni zbiór danych użyty do trenowania i testowania metody. W skład zbioru danych wchodziły sekwencje białkowe znajdujące się w bazie DisProt (wersja 3.6) oraz sekwencje z eksperymentu CASP7. Dodatkowo ze względu na ograniczenia metod składowych usunięto sekwencje dłuższe niż 1000 aminokwasów. W efekcie otrzymano zbiór testowy, który składał się z 566 białek i zawierał 232664 reszt aminokwasowych, spośród których 23,45% było oznaczonych jako wewnętrznie nieuporządkowane.

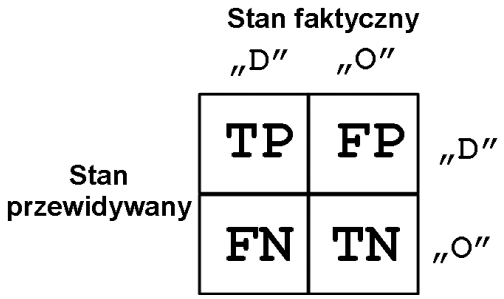
Następny zbiór danych, nazwany pdbRemark465, zbudowany został w oparciu o sekwencje struktur znajdujące się w bazie PDB. W celu usunięcia nadmiarowości danych wykorzystany został serwis internetowy PISCES. Sekwencje musiały spełniać następujące kryteria: pochodzić ze struktur otrzymanych metodą krystalografii rentgenowskiej, mieć rozdzielczość poniżej 2 Å, współczynnik R-factor poniżej 0,2. Ponadto długość sekwencji musiała znajdować się w zakresie 50-1000 aminokwasów, a poziom identyczności sekwencji do innych białek w zestawie musiał być mniejszy niż 20%. W efekcie otrzymano zbiór testowy

składający się z 1147 sekwencji (w sumie 289008 aminokwasów spośród których 6,28% oznaczone było znacznikiem REMARK465 w odpowiadających im plikach PDB).
W ostatecznej wersji program korzystał ze zbioru łączącego zbiory CASP7, DisProt i pdbRemark465. Dane te są dostępne do pobrania ze strony serwisu oraz zostały dołączone jako załącznik do publikacji opisującej metodę MetaDisorder (Kozlowski i Bujnicki, 2012).

6.2.1.3. Miary jakości przewidywania użyte w czasie uczenia i testowania metod

Dla każdej z reszt aminokwasowychw program zwraca przewidywanie, czy uznawana jest za zlokalizowaną w regionie wewnętrznie nieuporządkowanym (oznaczony dalej w tekście jako „D" – ang. *disorder*), czy w regionie ustrukturalizowanym („O" – ang. *order)*. W związku z tym możemy wydzielić cztery kategorie wyników:

- wyniki prawdziwie dodatnie (TP, ang. *true positives*; reszty aminokwasowe przewidziane jako należące do regionu wewnętrznie nieuporządkowanego znajdują się w takim regionie),
- wyniki prawdziwie ujemne (TN, ang. *true negatives*; reszty aminokwasowe należące do regionu ustrukturalizowanego są przewidziane jako takie),
- wyniki fałszywie dodatnie (FP, ang. *false positives*; reszty aminokwasowe, które błędnie sklasyfikowano jako wewnętrznie nieuporządkowane)
- wyniki fałszywie ujemne (FN, ang. *false negatives*; reszty aminokwasowe „błędnie" sklasyfikowane jako znajdujące się poza regionami nieuporządkowanymi) – ryc. 16.



Ryc. 16. Kategorie wyników przyjęte do oceny przewidywania wewnętrznego niuporządkowania białek. „D" – ang. *disorder* – region wewnętrznie nieuporządkowany, „O" – ang. *order* – region uporządkowany.

Jako pierwszej miary użyto współczynnika korelacji Matthewsa (Matthews, 1975) zdefiniowanego według następującego wzoru:

$$MCC = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

Miary MCC użyto jedynie do oceny poprawności przewidywań, nie była ona wykorzystana podczas uczenia metody.

Następną użytą miarą był współczynnik S_w (Jin i Dunbrack, 2005), wyrażony wzorem:

$$Sw = \frac{WdisorderTP - WorderFP + WorderTN - WdisorderFN}{Wdisorder(TN + FN) + Worder(TN + FP)}$$

gdzie $W_{disorder}$ to odsetek reszt aminokwasowych, które znajdują się w regionach uporządkowanych, a W_{order} to to odsetek odpowiadający ilości reszt aminokwasowych zlokalizowanych w regionach nieuporządkowanych.

Wzór na współczynnik S_w zawiera wagi, których zadaniem jest uniknięcie nadmiernego przewidywania klasy występującej częściej. W przeciwnym przypadku istnieje niebezpieczeństwo, że naiwny program, który przewidywałby wszystkie miejsca jako znajdujące się poza regionem nieuporządkowanym (klasa dominująca, która w zbiorze testowym pdbRemark465 stanowi 93,72% przypadków), byłby lepszy niż metoda, która popełniałaby pewną ilość błędów, ale ogólnie zwracałaby przewidywania dotyczące obu klas. Miara S_w przyjmuje wartości z zakresu [-1, 1], gdzie 0 oznacza przewidywania nie lepsze niż generowane losowo. Miara ta była głównym kryterium oceny jakości przewidywań poszczególnych metod składowych i samej meta-metody.

Część programów zwracając wynik jest w stanie ocenić, jak bardzo wiarygodne jest przewidywanie dla konkretnej reszty aminokwasowej. Odbywa się to poprzez zwrócenie wartości prawdopodobieństwa, gdzie wartości powyżej lub poniżej określonej granicy klasyfikowane są jako należące do określonej klasy (zwyczajowo przyjmuje się przedział [0, 1] i punkt decyzyjny w połowie przedziału; oznacza to, że w naszym przypadku program uznaje za wewnętrznie nieuporządkowane reszty aminokwasowe, dla których wartość prawdopodobieństwa jest większa lub równa 0,5). Dla programów tego typu możliwe jest użycie bardziej dokładnej miary do oceny wyników, jaką jest krzywa ROC (ang. *receiver operating characteristic*), a konkretnie pole powierzchni pod krzywą ROC (ang. *area under curve*). Krzywa ROC tworzona jest poprzez nałożenie na wykres czułości określanej także skrótem TPR (ang. *true positive rate*) i specyficzności określanej czasem skrótem TNR (ang. *true negative*

rate) w zakresach [0, 1] z określoną dokładnością. Miary TPR i TNR definiowane są w następujący sposób:

$$TPR = \frac{TP}{TP+FN} \qquad \text{TNR} = \frac{TN}{TN+FP}$$

W praktyce procedura wykreślania krzywej ROC polega na naniesieniu na wykres określonej liczby punktów odpowiadających czułości i specyficzności w miarę zmieniania progu decyzyjnego powyżej lub poniżej którego klasyfikujemy przypadki jako TN, TP, FN i FP (Robin et al., 2011). Pole powierzchni pod krzywą ROC można obliczyć wykorzystując metodę trapezów (Press et al., 2007). Miara AUC przyjmuje wartości z przedziału [0, 1], gdzie wartość 0,5 oznacza wynik dla klasyfikatora losowego.

6.2.1.4. Konstrukcja i testowanie metody

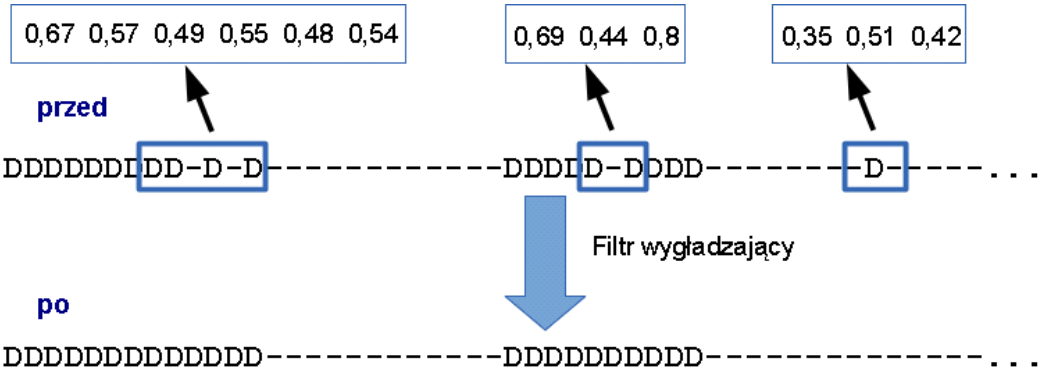
Zarówno do oceny metod składowych, jak i w czasie konstruowania meta-metody niezbędna jest ocena poprawności działania danego klasyfikatora na określonym zbiorze testowym. Oznacza to, że dla wszystkich 13 metod wchodzących w skład programu MetaDisorder konieczne było otrzymanie wyników dla wszystkich sekwencji ze zbiorów pdbRemark465, DISPROT i CASP7 (dane te są dostępne do pobrania ze strony serwisu http://iimcb.genesilico.pl/metadisorder/ oraz zostały dołączone jako załącznik do publikacji opisującej serwis MetaDisorder (Kozlowski i Bujnicki, 2012)).

Dla otrzymanych wyników można określić jakość danej metody poprzez wyznaczenie wartości TP, TN, FP i FN i zastosowanie wymienionych w poprzednim podrozdziale miar. Statystyczna istotność wyników została potwierdzona za pomocą testu *bootstrap*, który polegał w tym przypadku na 1000-krotnym losowym wyborze 80% sekwencji ze zbioru testowego i wyliczeniu średniej z próby (Carpenter i Bithell, 2000). W przypadku krzywych ROC dodatkowo użyto testu Wilcoxona (Rosner et al., 2006). Wśród testowanych metod najlepszy okazał się program iPDA, PrDOS oraz DISPROT – w zależności od użytej miary oceny (tabela 6.2.1.3.).

W najprostszym przypadku za meta-metodę można uznać prosty klasyfikator, który przeprowadza uśrednianie arytmetyczne. Oczywiście nie jest to najlepsze z możliwych rozwiązań. Mając dostęp do statystyk odzwierciedlających jakość metody, można zbudować inny wariant konsensusu, gdzie poszczególnym metodom składowym przypisywane są

odpowiednie wagi, niejako nagradzając lepsze metody. W ten sposób w ramach niniejszego projektu skonstruowane zostały dwa warianty oznaczone jako BinCons i FloatCons. Różnica między nimi polega na tym, że w przypadku wariantu BinCons użyto jedynie wartości binarnych („D" lub „O"), natomiast w przypadku FloatCons brane pod uwagę są ciągłe wartości prawdopodobieństwa (o ile program takie zwraca). Jako wag użyto wartości Sw.

Dodatkowo w BinCons i FloatCons zastosowano dwa zabiegi mające na celu poprawę jakości przewidywań. Po pierwsze, z przeprowadzonych badań (Xue et al., 2010) wynika, że regiony nieuporządkowane są nadmiernie reprezentowane w okolicach końców sekwencji białka. Z tego względu zastosowano prostą poprawkę statystyczną dla 15 końcowych aminokwasów, zwiększając prawdopodobieństwo klasyfikowania ich jako wewnętrznie nieporządkowane proporcjonalnie do wartości obserwowanych w zbiorze uczącym. Drugim zabiegiem poprawiającym jakość przewidywania jest użycie filtra wygładzającego (tzw. filtr dolnoprzepustowy), który ma za zadanie usuwać zbyt krótkie fragmenty przewidziane jako wewnętrznie nieuporządkowane lub naprawiać fragmenty nieciągłe zawierające kilkuaminokwasowe przerwy. Regiony wewnętrznie nieuporządkowane zbudowane są z sąsiadujących ze sobą reszt aminokwasowych, które tworzą różnej długości ciągłe segmenty nie zawierające krótkich przerw. Niestety podczas uśredniania wyników z wielu metod dość często dochodzi do sytuacji, gdzie pojedyncze reszty aminokwasowe klasyfikowane są nieprawidłowo, ponieważ nie uwzględniono sąsiedztwa innych reszt, a ich wartości prawdopodobieństwa są blisko progu decyzyjnego (ryc. 17).



Ryc. 17. Przykład działania filtra wygładzającego zastosowanego dla polepszenia wyników znajdujących się w okolicach progu decyzyjnego wynoszącego 0,5.

Prócz wyżej wymienionych wariantów programu MetaDisorder, zaprojektowano także algorytm meta-metody oparty na jednokierunkowej sieci neuronowej (ang. *feedforward neural network*). Dane wejściowe reprezentowane są przez wektor o długości 40 elementów. 21 elementów składowych określa typ aminokwasu (20 standardowych aminokwasów oraz „X" dla niestandardowych aminokwasów); 6 elementów reprezentuje strukturę drugorzędową przewidzianą za pomocą programów ACCpro i PSIPRED (po trzy elementy odzwierciedlające prawdopodobieństwo klasyfikacji reszty aminokwasowej jako należącego do helisy α, wstęgi β lub innego typu struktury drugorzędowej) oraz przewidywanie braku nieuporządkowania według 13 wcześniej użytych programów. Sieć neuronowa złożona jest z jednej warstw ukrytej zbudowanej z 40 neuronów, a proces uczenia przebiegał przez 30 tysięcy iteracji.

W przypadku trenowania dowolnej metody wliczając w to meta-metody niezbędne jest podzielenie zbioru danych na zbiór uczący i zbiór testowy, tak, aby uniknąć przetrenowania metody. Zwykle robi się to za pomocą n-krotnego sprawdzianu krzyżowego (ang. *n-fold crossvalidation*), która polega na tym, że zbiór danych dzieli się losowo na np. 10 części (mówimy wtedy o tzw. 10 krotnej walidacji krzyżowej) i 9 z nich użyte zostaje w czasie trenowania metody, a jedna część służy do przeprowadzenia testów. Następnie procedurę powtarza się n razy, wymieniając część, na której testujemy metodę. W ten sposób zwiększamy szanse, że nasz klasyfikator będzie w stanie przewidzieć określoną cechę także dla nowych danych, nieużytych do trenowania.

W tabeli 4 zamieszczono ostateczny ranking skuteczności opisanych wyżej wersji programu MetaDisorder i programów składowych, do których został porównany. Jak widać z wyników, poszczególne wersje programu MetaDisorder wypadają lepiej bez względu na użyte kryterium oceny. Ponadto zbadano istotność otrzymanych wyników, wykorzystując test Wilcoxona (tabela 5).

Tabela 5. Wyniki dwustronnego testu Wilcoxona dla wartości AUC na łączonym zbiorze testowym (pdbRemark465, DISPROT i CASP7).

	Float Cons	Bin Cons	VSL2	Dis PSSMP	iPDA	IUPred short	PrDOS	DISO PRED	IUPred long	POO DLE-L	RONN	POO DLE-S
FloatCons	x											
BinCons	0,005	x										
VSL2	0,005	0,007	x									
DisPSSMP	0,005	0,005	0,005	x								
iPDA	0,005	0,005	**0,241**	0,007	x							
IUPred short	0,005	0,005	0,005	**0,093**	0,005	x						
PrDOS	0,005	0,005	0,022	0,017	0,005	0,005	x					
DISOPRED	0,005	0,005	0,005	**0,445**	0,005	**0,028**	0,005	x				
IUPred long	0,005	0,005	0,005	**0,203**	0,005	**0,878**	0,005	**0,114**	x			
POODLE-L	0,005	0,005	0,005	**0,721**	0,005	0,009	0,005	**0,386**	0,013	x		
RONN	0,005	0,005	0,005	**0,169**	0,005	0,005	0,005	**0,028**	0,005	0,007	x	
POODLE-S	0,005	0,005	0,005	0,005	0,005	0,005	0,005	0,005	0,005	0,005	0,005	x

Wartości pogrubione oznaczają pary metod, pomiędzy którymi różnica między wartościami AUC jest statystycznie nieistotna na poziomie wartości *p* wynoszącym 0,05.

Tabela 4. Ranking skuteczności różnych wersji programu MetaDisorder na łączonym zborze testowym (pdbRemark465, DISPROT i CASP7). Podano wartości średnie wraz z odchyleniem standardowym obliczonym za pomocą testu *bootstrap*. Czcionka pogrubioną zaznaczono najlepsze wyniki (osobno uwzględniając metody składowe oraz różne wersje programu MetaDisorder).

Metoda	**Sw**	**MCC**	**AUC**
ANNCons	**0,609 ± 0,010**	0,455 ± 0,005	0,863 ± 0,003
FloatCons	0,608 ± 0,007	0,475 ± 0,008	**0,868 ± 0,002**
BinCons	0,599 ± 0,007	**0,487 ± 0,008**	0,843 ± 0,003
iPDA	***0,555 ± 0,006***	0,419 ± 0,006	0,829 ± 0,004
DISPROT(vls2)	0,539 ± 0,005	0,399 ± 0,005	**0,830 ± 0,001**
DISOPRED	0,481 ± 0,006	0,436 ± 0,006	0,778 ± 0,003
POODLE-S	0,474 ± 0,009	0,423 ± 0,010	0,828 ± 0,004
PrDOS	0,469 ± 0,007	**0,442 ± 0,008**	0,810 ± 0,006
POODLE-L	0,464 ± 0,010	0,397 ± 0,010	0,794 ± 0,004
RONN	0,450 ± 0,006	0,350 ± 0,007	0,762 ± 0,006
IUPred (long)	0,445 ± 0,006	0,412 ± 0,007	0,788 ± 0,002
DisPSSMP	0,442 ± 0,012	0,377 ± 0,012	0,776 ± 0,004
IUPred (long)	0,432 ± 0,008	0,392 ± 0,009	0,787 ± 0,004
Spritz (long)	0,418 ± 0,009	0,377 ± 0,010	-
Pdisorder	0,383 ± 0,007	0,350 ± 0,007	-
DISpro	0,355 ± 0,006	0,411 ± 0,008	-
Spritz (short)	0,334 ± 0,007	0,306 ± 0,007	-
DisEMBL	0,289 ± 0,007	0,232 ± 0,006	-
GlobPlot	0,187 ± 0,004	0,172 ± 0,004	-

6.2.1.5. Skuteczność metody w eksperymencie CASP8

Eksperyment CASP (ang. ***C**ritical **A**ssessment of techniques for protein **S**tructure **P**rediction)* to międzynarodowy konkurs, który odbywa się co 2 lata (w 2012 roku miała miejsce 10. edycja eksperymentu) i ma na celu ocenę zdolności ludzi i programów do przewidywania struktury trzeciorzędowej oraz innych właściwości białek (takich jak wewnętrzne nieuporządkowanie czy funkcja białka). Jest to test ślepy, ponieważ przeprowadzany jest on na sekwencjach z białek, których struktura nie jest jeszcze publicznie znana (udało się otrzymać strukturę metodami doświadczalnymi, jednak autorzy nie zdeponowali jeszcze danych np. w bazie PDB). W eksperymencie mogą brać udział zarówno tzw. „grupy ludzkie" (na wysłanie wyników mają one 2 tygodnie) oraz „grupy serwerowe" (programy mają 3 dni na odesłanie przewidywań). Jest to konkurs otwarty oraz anonimowy (w czasie rejestracji grupie przydzielany jest losowy numer). Aby uniknąć jakiegokolwiek oszustwa lub nieświadomego błędu ludzkiego, także eksperci oceniający wyniki nie wiedzą, kto stoi za jaką grupą. Dopiero po zakończeniu eksperymentu i wyliczeniu wszystkich statystyk, nazwy serwisów są łączone z numerami grup.

W celu niezależnego przetestowania programu MetaDisorder, opisane wyżej trzy wersje programu zostały zrejestrowane w konkursie CASP8 i otrzymały następujące numery:

- BinCons numer 153
- FloatCons numer 297
- ANNCons numer 133

W ósmej edycji eksperymentu CASP brało udział w sumie 25 grup, które postawiły sobie za cel przewidywanie braku uporządkowania. W tabeli 6 pokazano 10 najlepszych grup, wśród których znalazły się wszystkie trzy wersje programu MetaDisorder. FloatCons i BinCons były najlepsze biorąc pod uwagę miarę S_w. Także uwzględniając inne miary, MetaDisorder znajduje się w czołówce rankingu. Co ciekawe, wynik ANNCons jest słabszy niż można by się spodziewać po wynikach testów na wewnętrznym zbiorze łączącym pdbRemark465, DISPROT (patrz tabela 4). Podobnie pewnym zaskoczeniem może być niewielka różnica między FloatCons a BinCons. Wydawałoby się, że tworząc konsensus z wartości odzwierciedlających prawdopodobieństwo bycia w regionie wewnętrznie nieuporządkowanym, a nie samą informację

binarna „D” lub „O” powinniśmy otrzymać lepszy wynik tak się nie stało (różnica między obiema metodami jest statystycznie nieistotna).

Tabela 6. Wyniki eksperymentu CASP8 w kategorii przewidywanie braku nieuporządkowania. Z 25 grup pokazano 10 najlepszych, tabela jest posortowana po wartości współczynnika S_w. Na podstawie oficjalnych wyników opublikowanych w (Noivirt-Brik et al., 2009) przez niezależny zespół ekspertów odpowiedzialnych za ocenę przewidywań. Pogrubioną czcionką zaznaczono wyniki dla różnych wersji programu MetaDisorder, natomiast na czerwono zaznaczono najlepsze wyniki.

Numer grupy	Czułość	Specyficzność	S_w	AUC
153 (BinCons)	**0.758 ± 0.048**	**0.904 ± 0.004**	**0.662 ± 0.048**	**0.9078 ± 0.0186**
297 (FloatCons)	**0.741 ± 0.050**	**0.920 ± 0.003**	**0.661 ± 0.050**	**0.8967 ± 0.0208**
69	0.796 ± 0.039	0.864 ± 0.004	0.660 ± 0.039	0.8958 ± 0.0185
97	0.727 ± 0.047	0.917 ± 0.004	0.644 ± 0.047	0.9083 ± 0.0176
379	0.706 ± 0.053	0.938 ± 0.004	0.644 ± 0.053	0.9084 ± 0.0158
450	0.694 ± 0.040	0.927 ± 0.003	0.621 ± 0.040	0.8857 ± 0.0159
453	0.641 ± 0.061	0.978 ± 0.001	0.619 ± 0.061	0.9184 ± 0.0152
229	0.657 ± 0.049	0.955 ± 0.003	0.612 ± 0.049	0.9052 ± 0.0180
133 (ANNCons)	**0.711 ± 0.044**	**0.894 ± 0.004**	**0.605 ± 0.044**	**0.8295 ± 0.0245**
161	0.646 ± 0.066	0.942 ± 0.004	0.588 ± 0.066	0.8953 ± 0.0206

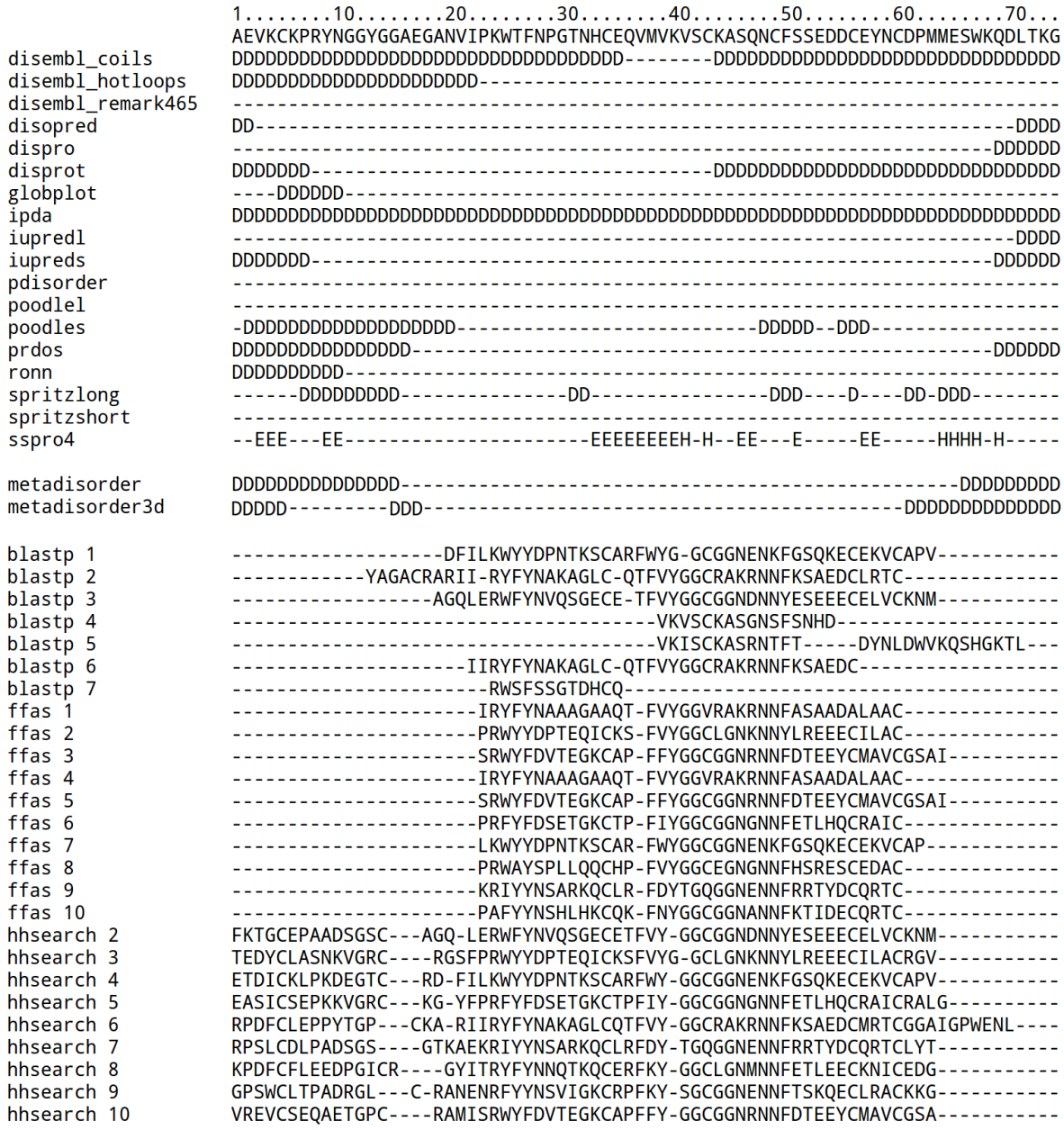
6.2.2. MetaDisorder3D – meta-metoda oparta na występowaniu przerw w przyrównaniach sekwencyjnych najbliższych homologów

Analizując wyniki z metaserwera GeneSilico do wykrywania zwojów, łatwo można zauważyć, że zestawiając ze sobą wyniki programów do przewidywania wewnętrznego nieuporządkowania z przyrównaniami generowanymi przez programy do wykrywania zwoju, regiony przewidziane jako wewnętrznie nieuporządkowane w znacznym stopniu pokrywają się z przerwami w przyrównaniach (ryc. 18). W oczywisty sposób nasuwa się kilka pytań. Po pierwsze, jak duża jest korelacja między obiema cechami? Po drugie, czy dodanie tej informacji do konsensusu przyniosłoby poprawę i jeśli tak, to jak to zrobić?

Aby odpowiedzieć na pierwsze z pytań, skonstruowano program oparty tylko i wyłącznie na przyrównaniach tworzonych przez następujące programy: PSI-BLAST (z włączonym i wyłączonym filtrowaniem regionów o niskiej złożoności sekwencyjnej, oznaczone odpowiednio jako pdblast i blastp), FFAS, HHsearch (przeszukiwanie baz PDB70 i CDD), GenTHREADER, Phyre oraz PCONS. Choć pozornie problem wydaje się łatwy do rozwiązania, na tym etapie powstaje zasadnicze pytanie: jaką wagę nadać poszczególnemu przyrównaniu w zależności od rodzaju metody oraz miary jakości? Ponieważ każdy z programów używa własnego systemu oceniającego, przyjęto podział przyrównań na dobre, średnie i słabe zgodnie z progami odcięcia

z tabeli 6.1.2.1. Ponadto mamy 8 typów programów (jeśli wersje pdblast, blastp oraz HHsearch przeszukujący inne bazy uznamy za oddzielny typ). W ten sposób nasz problem możemy zredukować do określenia 24 wag (8 programów x 3 poziomy jakości przyrównań). Dodatkowo w celu dalszego zredukowania przestrzeni potencjalnych rozwiązań przyjęto, że poszukiwane wagi mogą przyjmować wartości dyskretne ze skokiem wynoszącym 1 z przedziału od 1 do 10. Ostatecznie otrzymujemy 10^{24} możliwości, więc niezbędne jest użycie algorytmu lub heurystyki efektywnie próbkującego przestrzeń alternatywnych rozwiązań. W tym celu wykorzystano algorytm genetyczny zaimplementowany za pomocą biblioteki PyEvolve.

Jako zbioru uczącego użyto danych z eksperymentu CASP8 dostępnych w bazie metaserwera GeneSilico. Należy podkreślić, że zestaw ten może posiadać sekwencje homologiczne do znanych białek, oraz to, że nie reprezentuje on w równomierny sposób całej przestrzeni znanych zwojów białkowych. Jednak z drugiej strony korzystając z przewidywań, które zgromadzone są w bazie danych metaserwera GeneSilico, można łatwo otrzymać zbiór danych, który zawiera wszelkie niezbędne informacje, a ponadto, do którego można odnieść się ze względów historycznych (na sekwencjach z eksperymentu CASP testowana była większość istniejących programów), więc możliwe jest porównanie do innych metod. Oczywiście korzystając z takiego zestawu danych, ciągle trzeba wykazać istotność statystyczną wyników i stosować techniki zapobiegające przeuczeniu metody. Podobnie jak w poprzednich wersjach programu MetaDisorder, także tu zastosowano test *bootstrap* oraz 10-krotny sprawdzian krzyżowy.



Ryc. 18. Porównanie wyników programów do przewidywania wewnętrznego nieuporządkowania i przyrównań z programów do wykrywania zwoju. Dodatkowo pokazano wynik programu FloatCons (oznaczonego jako „metadisorder") oraz wynik programu MetaDisorder3D.

Ponieważ program MetaDisorder3D opierał się na danych pochodzących z CASP8, oczywistym było, że w pełni niezależny test powinien zostać przeprowadzony w czasie edycji CASP9. Ponadto MetaDisorder3D został stworzony z myślą o integracji z wersją FloatCons w celu sprawdzenia, czy taki zabieg przyniesie korzyść. Dlatego też statystyki dotyczące programu MetaDisorder3D zostaną zaprezentowane w następnym podrozdziale omawiającym wersję łączącą wspomniane programy.

6.2.3. MetaDisorderMD = MetaDisorder + MetaDisorder3D

Program MetaDisorderMD stanowi połączenie programu FloatCons z programem MetaDisorder3D. Jego celem było pokazanie, że dodanie informacji pochodzącej z przyrównań wygenerowanych przez programy do wykrywania zwoju może poprawić wyniki przewidywania wewnętrznego nieuporządkowania białek. Podobnie jak w przypadku konstruowania programu MetaDisorder3D pojawia się pytanie, w jaki sposób optymalnie połączyć oba komponenty. W tym celu wykorzystano podobny algorytm genetyczny jako poprzednio z tym wyjątkiem, że do wektora reprezentującego MetaDisorder3D dodano drugi wymiar reprezentujący wagi dla FloatCons i MetaDisorder3D. W ten sposób algorytm genetyczny miał za zadanie optymalizowanie następującej dwuwymiarowej tablicy:

[[FR1, FR1, FR1, …, FR8, FR8, FR8] [FloatCons, MetaDisorder3D]

gdzie FR oznacza wagę dla przyrównania dobrej jakości (kolor zielony), średniej jakości (kolor pomarańczowy) i słabej jakości (kolor czerwony) pochodzącego z jednej z ośmiu metod do wykrywania zwoju białka.

Podobnie jak w przypadku uczenia i testowania programu MetaDisorder3D, także w tym przypadku, wykorzystano zbiór testowy oparty na sekwencjach pochodzących z eksperymentu CASP8 i zastosowano tą samą procedurę konstrukcji programu oraz oceny istotności wyników (test *bootstrap* oraz 10-krotny sprawdzian krzyżowy).

6.2.4. Program MetaDisorderMD2 a miara S_{ww}

W przypadku programu MetaDisorderMD algorytm genetyczny miał za zadanie zoptymalizować wagi pod kątem miary S_w jednak, jeśli weźmiemy pod uwagę wyniki przedstawione w tabeli 6 możemy zauważyć, że uczenie metody pod kątem jednej miary nie zawsze skutkuje dobrym wynikiem pod względem innej. Dodatkowo, można zastanawiać się czy aktualnie stosowane kryteria oceny są idealne i czy nie da się skonstruować lepszych miar, które lepiej sprawdzają się w ocenie danego problemu. Program MetaDisorderMD2 jest próbą odpowiedzi na wszystkie te wątpliwości. Jego architektura, sposób tworzenia jest identyczny jak w przypadku MetaDisorderMD, jedyna różnica polega na funkcji oceny, która optymalizuje wagi pod kątem innej miary. Nową miarę nazwano S_{ww}.

Miara S_{ww} jest swego rodzaju połączeniem miary AUC z miarą S_w. Miara ta liczona jest ze wzoru na S_w, jednak dodatkowo zmieniany jest próg decyzyjny tak jak to ma miejsce w

przypadku konstruowania krzywej ROC (w zakresie [0, 1] przy kroku 0,01). Następnie wszystkie wartości S_w są sumowane i dzielone przez 100.

Aby ocenić skuteczność zastosowania miary S_{ww} program MetaDisorderMD2 został przetestowany w ramach eksperymentu CASP9.

6.2.5. Wyniki programu MetaDisorder w czasie eksperymentu CASP9

Wszystkie opisane wyżej warianty programów do przewidywania wewnętrznego nieuporządkowania były uczone na zbiorze sekwencji z eksperymentu CASP8 (tabela 7). Ponadto programy te poddano niezależnej ocenie w czasie eksperymentu CASP9 (tabela 8). Oba testy potwierdziły, że dodanie przyrównań pochodzących z metod do rozpoznawania zwoju poprawia jakość przewidywania regionów wewnętrznie nieuporządkowanych. Wyniki programów MetaDisorderMD i MetaDisorderMD2 były najlepsze. Jedynie pod względem miary AUC program PrDOS był lepszy o 1.4% niż MetaDisorderMD2. Wynik programu MetaDisorder3D jednoznacznie wskazuje, że przyrównania z programów do rozpoznawania zwoju niosą dużo informacji na temat wewnętrznego nieuporządkowania, jednak cecha ta nie jest wystarczająca do stworzenia dobrego klasyfikatora. Dopiero w połączeniu z programem FloatCons otrzymujemy poprawę skuteczności.

Tabela 7. Wyniki testów na zbiorze CASP8 dla programów FloatCons, MetaDisorder3D, MetaDisorderMD, MetaDisorderMD2. Pogrubioną czcionką zaznaczono najlepsze wyniki dla każdej z miar.

Metoda	MCC	Sw	AUC
FloatCons	**0.654 ± 0.041**	0.606 ± 0.023	0.904 ± 0.009
MetaDisorder3D	0.589 ± 0.047	0.519 ± 0.024	0.833 ± 0.014
MetaDisorderMD	0.558 ± 0.034	0.684 ± 0.023	0.927 ± 0.011
MetaDisorderMD2	0.607 ± 0.042	**0.684 ± 0.022**	**0.929 ± 0.017**

Tabela 8. Wyniki eksperymentu CASP9 w kategorii przewidywania regionów wewnętrznie nieuporządkowanych. W tabeli umieszczono cztery najlepsze metody (górny panel) oraz dodatkowo do celów porównawczych wynik metody FloatCons i MetaDisorder3D (dolny panel). Tabela zawiera wybór danych z (Monastyrskyy et al., 2011). Pogrubioną czcionką zaznaczono najlepsze wyniki dla każdej z miar.

Numer grupy	Czułość	Specyficzność	S_w	AUC
MetaDisorderMD	**0.654 ± 0.012**	0.821 ± 0.010	0.476 ± 0.006	0.818 ± 0.008
MetaDisorderMD2	**0.653 ± 0.013**	0.860 ± 0.012	**0.516 ± 0.010**	0.841 ± 0.014
PrDOS2	0.609 ± 0.008	0.857 ± 0.003	0.509 ± 0.002	**0.855 ± 0.010**
MULTICOM	0.651 ± 0.003	0.851 ± 0.004	0.500 ± 0.003	0.821 ± 0.008
MetaDisorder3D	0.574 ± 0.020	0.854 ± 0.009	0.427 ± 0.009	0.795 ± 0.011
FloatCons	0.411 ± 0.016	**0.948 ± 0.008**	0.391 ± 0.007	0.784 ± 0.012

6.3. Przewidywanie domen w białkach

Problem podziału białek na domeny jest szczególnie istotny, ponieważ stanowią one niezależne moduły białek o dużej autonomii zarówno, jeśli chodzi o proces zwijania jak i ich funkcję (Levitt i Chothia, 1976; Porter i Rose, 2012). Istnieje kilkanaście programów komputerowych, które przewidują lokalizację domen w białku (duża część z nich została włączona do metaserwera GeneSilico (patrz tabela 2), jednak ich skuteczność ciągle jest niewystarczająca. W niniejszym rozdziale opisany zostanie program nazwany DomainSVM, który ma za zadanie poprawić tą sytuację. Program ten zbudowano opierając się na metodzie uczenia maszynowego, a konkretnie wykorzystano tzw. maszynę wektorów nośnych. Program przetestowano na zbiorze testowym, a następnie wyniki jego działania porównano do siedmiu istniejących programów do przewidywania domen w białkach.

6.3.1. Definicja domeny i jej granic

Problem przewidywania domen można sprowadzić do identyfikacji ich granic. Otrzymujemy dwie klasy: reszty aminokwasowe budujące domeny oraz reszty aminokwasowe klasyfikowane jako graniczne. W tym momencie pojawiają się dwa problemy. Po pierwsze, zwykle poprzez granicę białka rozumiemy region w którym jedna domena przechodzi w drugą czyli granica odnosi się do relatywnie krótkiego fragmentu sekwencji (w skrajnym przypadku jest to pojedyncza reszta aminokwasowa, czasem wydłużona o reszty budujące pętle). Jeśli chcielibyśmy przewidzieć tak zdefiniowaną granicę domeny, regiony znajdujące się wewnątrz domen stanowiłyby ponad 99% reszt aminokwasowych dowolnego zbioru testowego. Takie dysproporcje na pewno odbiłyby się na skuteczności klasyfikatora. Po drugie, taka granica domeny jest dużym uproszczeniem, ponieważ jest zero-jedynkowa i nie bierze ona pod uwagę tego jak blisko prawdziwej granicy znajduje się przewidywanie. Oczywistym jest, że jeśli klasyfikator zwróci wynik oddalony od prawdziwej granicy o 3 lub 5 reszt aminokwasowych to jest to lepszy wynik, niż gdy przewidywany wynik będzie oddalony o 30-50 reszt. Aby ominąć oba te problemy w przypadku przewidywania domen stosuje się tzw. regułę ± 20 aminokwasów (George i Heringa, 2002). Według tej reguły wszystkie reszty aminokwasowe położone ± 20 aminokwasów od prawdziwej granicy traktowane są jako reszty graniczne.

Przykładowo załóżmy, że sekwencja białka wygląda w następujący sposób (jest to obraz uproszczony do celów poglądowych, w praktyce dysproporcje są jeszcze większe):

```
AAAAAAAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBB---CCCCCCCCCCCCCC...
-----------------1-------------------------1111------------...
```

gdzie kolejne litery oznaczają poszczególne domeny, a znak „-” w pierwszym przypadku oznacza region łącznikowy, a w drugim oznacza obszar wewnątrz domeny. Stosując regułę ± 5 aminokwasów otrzymujemy (jak wspomniano powyżej w praktyce stosujemy próg ± 20 aminokwasów):

```
------------1111111111---------------1111111111111--------...
```

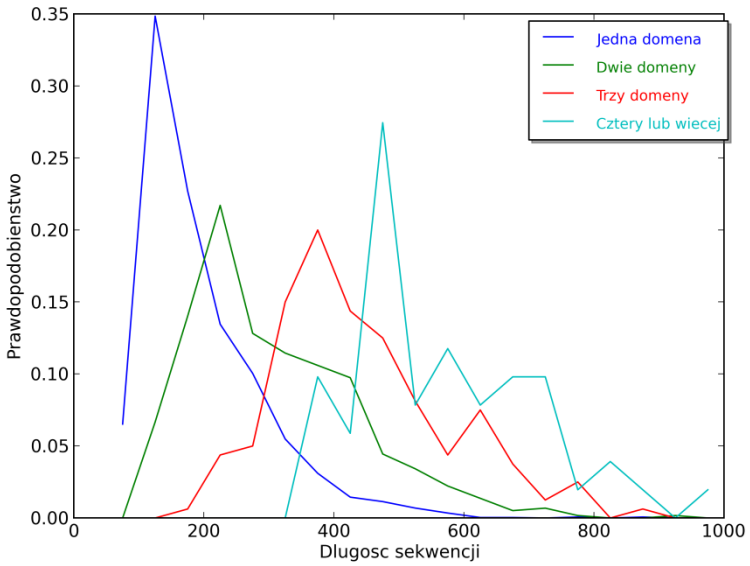
W ten sposób rozciągamy mniej liczną klasę. Dodatkowo zabieg ten ma tą zaletę, że za wytworzenie granicy w białku odpowiada lokalne otoczenie „granicy” a nie pojedyncza reszta aminokwasowa, a więc stosując regułę ± 20 aminokwasów nasz klasyfikator otrzymuje więcej informacji.

Należy zwrócić uwagę, że duża część białek posiada jedynie jedną domenę, a liczba domen w dużym stopniu zależy od długości sekwencji białka (Wheelan et al., 2000). Im dłuższe białko tym większe prawdopodobieństwo, że będzie ono posiadać więcej niż jedną domenę. W ten sposób nasz problem można podzielić na dwa mniejsze. Po pierwsze chcemy wiedzieć czy białko zawiera jedną czy więcej domen (jest to zadanie względnie łatwe), a jeśli tak jest, to gdzie dokładnie znajdują się granice poszczególnych domen (to zadanie jest już znacznie trudniejsze do rozwiązania). W ten sposób możemy skonstruować program złożony z dwóch warstw. W pierwszej warstwie program będzie miał na celu przewidzieć czy białko zbudowane jest z jednej domeny czy z wielu. Następnie, jeśli sekwencja zostanie oznaczona jako posiadająca więcej niż jedną domenę uruchomiona zostanie druga warstwa, której zadaniem podzielenie sekwencji na odcinki wewnątrz domenowe oraz odcinki graniczne (zgodnie z regułą ± 20 aminokwasów).

6.3.2. Zbiór testowy

W celu zbudowania klasyfikatora i jego porównania do obecnie istniejących programów skonstruowano specjalny zbiór danych (nazwany 90_1000_CATH35). Opiera się on na sekwencjach pochodzących z bazy domen CATH (wersja 3.4), z której wybrano sekwencje spełniające następujące kryteria: długość 90-1000 reszt aminokwasowych oraz identyczność sekwencji na poziomie 35%. Dodatkowo, aby usunąć wątpliwe przypadki liczba domen w danej

sekwencji musi być zgodna z ilością domen w innej bazie domen tj. w bazie PFAM. Dodatkowo ze zbioru 90_1000_CATH35 wydzielono dwa mniejsze podzbiory (nazwane odpowiednio singleANDmulti i multiBoundary). Podsumowanie zbiorów testowych przedstawiono w tabeli 9. Biorąc pod wagę dystrybucję domen w zbiorze testowym (ryc. 19) oraz statystyki podane w tabeli można zauważyć następujące zależności. Zdecydowana większość sekwencji (75,7%) zawiera jedynie jedną domenę, a więc nie niesie informacji dotyczącej granic domen. Z tego powodu konstruując klasyfikator drugiej warstwy można ograniczyć się do sekwencji zawierających jedynie dwie lub więcej domen (zbiór multiBoundary). Ponadto zbiór 90_1000_CATH35 zawiera blisko trzy razy więcej białek jednodomenowych niż białek wielodomenowych. Aby zredukować tą dysproporcję spośród 2795 białek jednodomenowych losowo wybrano 899 tworząc zbiór singleANDmulti). Zbiór ten wykorzystano do zbudowania pierwszej warstwy, czyli klasyfikatora odróżniającego białka jednodomenowe od wielodomenowych.



Ryc. 19. Rozkład prawdopodobieństwa długości białka do liczby domen w białku dla zbioru testowego 90_1000_CATH35.

Tabela 9. Statystyki zbiorów wykorzystanych w czasie konstruowania program DomainSVM

	90_1000_CATH35	**singleANDmulti**	**multiBoundary**
Liczba białek	3694	1798	899
Liczba białek jednodomenowych	2795	899	0
Liczba białek dwudomenowych	630	630	630
Liczba białek zawierających trzy domeny	199	199	199
Liczba białek zawierających >4 domeny	70	70	70
Reszty aminokwasowe poza regionami granicznymi	786793 (93,77%)	426735 (89,09%)	255376 (83,01%)
Reszty aminokwasowe w regionach granicznych	52280 (6,23%)	52280 (10,91%)	52280 (16,99%)
Całkowita liczba reszt aminokwasowych	839073	479015	307656

6.3.3. Implementacja metody

Jak wspomniano wyżej program DomainSVM zbudowany został z dwóch oddzielnych klasyfikatorów. Pierwszy klasyfikator odpowiada za rozróżnienie białek jednodomenowych od wielodomenowych. W przypadku białek wielodomenowych program przechodzi do drugiej warstwy, która przewiduje lokalizacje granic domen. Oba klasyfikatory skonstruowano w używając maszyny wektorów nośnych, w której jako funkcji jądrowej (ang. *kernel*) użyto radialnej funkcji bazowej (ang. *radial basis function*, RBF). Choć maszyna SVM pozwala na użycie wielu innych funkcji jądrowych wykazano, że inne funkcje są szczególnym przypadkiem funkcji RBF. Niezbędna jest jedynie optymalizacja specjalnych parametrów C i γ, która zapewnia znalezienie najlepszego klasyfikatora (Keerthi i Lin, 2003; Lin HT, 2003).

6.3.4. Miary służące do oceny klasyfikatorów

W czasie uczenia i testowania metody zastosowano miary i techniki statystyczne opisane w rozdziale dotyczącym metod przewidujących wewnętrzne nieuporządkowanie. Różnice dotyczyły jedynie dwóch kwestii. Po pierwsze w czasie trenowania użyto 5-krotnego sprawdzianu krzyżowego zamiast 10-krotnego. Po drugie zamiast miary Sw użyto dokładności (ang. *accuracy*, ACC) zdefiniowanej w następujący sposób:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Należy podkreślić, że wartości TP, TN, FP i FN mają inne znaczenie w zależności od warstwy programu DomainSVM, która rozpatrujemy. W przypadku pierwszej warstwy rozróżniającej białka jednodomenowe od wielodomenowych wartości te maja następujące znaczenie:

- wynik TP oznacza poprawne sklasyfikowanie białka jednodomenowego,
- wynik FN oznacza poprawne sklasyfikowanie białka wielodomenowego,
- wynik FP oznacza błędne sklasyfikowanie białka jako jednodomenowego,
- wynik FN oznacza błędne sklasyfikowanie białka jako wielodomenowego.

W przypadku uczenia drugiej warstwy programu DomainSVM, która przewiduje obszary graniczne domen wartości te maja następujące znaczenie:

- wynik TP oznacza poprawne sklasyfikowanie obszary graniczne,
- wynik FN oznacza poprawne sklasyfikowanie obszary znajdujące się poza granicą białka,
- wynik FP oznacza błędne sklasyfikowanie obszary graniczne,
- wynik FN oznacza błędne sklasyfikowanie obszary znajdujące się poza granicą białka.

6.3.5. Przewidywanie liczby domen

Klasyfikator odróżniający białka jednodomenowe od wielodomenowych został wytrenowany na podstawie następujących cech:

- długość sekwencji,
- entropia Shannona i entropia metryczna (Shannon, 1948),
- średnia hydrofobowość GRAVY (ang. *grand average of hydropathy*) z wykorzystaniem skali hydrofobowości Kyte-Doolittle'a (Kyte i Doolittle, 1982),
- przewidywanie wewnętrznego nieuporządkowania według programu RONN wyrażone w formie procentowej zawartości regionów IUR, obecności regionu IUR o długości przynajmniej 10, 20, 30, 40, 50 i 60 reszt aminokwasowych, liczbie regionów IUR o długości powyżej 10 reszt,
- przewidywanie struktury drugorzędowej za pomocą programów PSIPRED i Sspro4,
- przewidywanie relatywnej dostępności dla rozpuszczalnika (RSA) według programu ACCpro dla trzech progów 0%, 5% i 25% dostępności
- informacja dotycząca liczby domen u najbliższych homologów (w pierwszej kolejności za pomocą programu HHsearch identyfikowane jest 10 najlepszych sekwencji homologicznych, następnie dla tych sekwencji pobierana jest informacja na temat liczby domen według bazy CATH).

Cechami, które można otrzymać bezpośrednio na podstawie sekwencji i które mają największy wpływ na jakość przewidywania są jej długość i entropia. Dopiero łącząc te cechy oraz uwzględniając liczbę domen w najbliższych sekwencjach homologicznych otrzymujemy najlepszy klasyfikator. Ostatecznie klasyfikator, który uwzględnia wszystkie te cechy ma 95% skuteczność (tabela 10). Jeśli pominiemy informację na temat liczby domen w białkach homologicznych skuteczność metody spada do 78,5%. W praktyce oznacza to, że klasyfikator ten ma akceptowalną moc dyskryminacji, także dla białek o nieznanym zwoju białka.

Tabela 10. Wpływ poszczególnych cech białka na rozróżnianie białek jednodomenowych od wielodomenowych. Parametry C i γ pokazane są jako potęgi o podstawie 2.

Cecha	**ACC**	**parameter C**	**parameter γ**
średnia hydrofobowość (GRAVY)	0,5547	5	5
relatywna dostępność dla rozpuszczalnika (RSA)	0,6232	13	-1
struktura drugorzędowa (ss)	0,6726	1	5
wewnętrzne nieuporządkowanie (IUP)	0,7000	1	1
entropia	0,7452	1	7
długość	0,7364	16	-21
długość, entropia, IUP, GRAVY, ss, RSA	0,7855	4	-5
długość, entropia, IUP, GRAVY, ss, RSA, homologia	0,9509	14	-14

W przypadku przewidywania dowolnej cechy istotne jest pokazanie jak nowy klasyfikator wypada na tle już istniejących. W tabeli 11 porównano wynik działania programu DomainSVM z innymi tego typu programami. Interpretując wyniki należy zwrócić uwagę na następujący fakt. Najlepsza metodą jest DomainSVM w wersji, która uwzględnia homologię, następny jest program FIEFdom, który także korzysta z informacji o domenach w białkach homologicznych. Metoda DomainSVM, ale bez uwzględnienia informacji dotyczącej homologów, jest już dużo słabsza, choć ciągle jej wynik jest lepszy od innych programów (poza wcześniej wspomnianym programem FIEFdom).

Tabela 11. Porównanie programu DomainSVM do innych programów. Statystyki dotyczą zbioru testowego singleANDmulti.

Program	TP	TN	FP	FN	Czułość	Specyficzność	ACC	MCC
GLOBPLOT	828	266	633	71	0,9210	0,5667	0,6085	0,2779
DOMpro	817	309	590	82	0,9088	0,5807	0,6263	0,3061
PPRODO	507	715	184	392	0,5640	0,7337	0,6796	0,3693
Scooby-domain (Domcut)	356	847	52	543	0,3960	0,8726	0,6691	0,4037
DoBo	746	550	349	153	0,8298	0,6813	0,7208	0,4525
Scooby-domain (Greedy)	625	703	196	286	0,6952	0,7613	0,7386	0,4790
DomPred (DOMSSEA)	694	684	215	205	0,7720	0,7635	0,7664	0,5329
DomPred (DPS)	694	687	212	205	0,7720	0,7660	0,7681	0,5362
DomainSVM*	671	735	164	228	0,7464	0,8036	0,7820	0,5654
FIEFdom	831	826	73	68	0,9244	0,9193	0,9216	0,8432
DomainSVM	845	865	34	54	0,9399	0,9613	0,9511	0,9023

* wersja która nie uwzględnia liczby domen w białkach homologicznych (długość, entropia, IUP, GRAVY, ss, RSA)

6.3.6. Przewidywanie granic domen

W przypadku, gdy pierwsza warstwa programu DomainSVM sklasyfikuje daną sekwencję jako pochodzącą z białka wielodomenowego uruchamiana jest druga warstwa, której zadaniem jest przewidzenie granicy domen na poziomie poszczególnych reszt aminokwasowych. Klasyfikator drugiej warstwy programu DomainSVM uwzględnia następujące cechy:

- rodzaj reszty aminokwasowej,
- przewidywany stan wewnętrznego nieuporządkowania (według programu RONN),
- przewidywana struktura drugorzędowa (według programów PSIPRED i Sspro),
- przewidywana dostępność dla rozpuszczalnika (według programu ACCpro dla trzech progów 0%, 5% i 25% dostępności),
- relatywna pozycja reszty aminokwasowej w sekwencji (pozycja reszty podzielona przez długość sekwencji),
- odległość od końca N i odległość od końca C,
- przewidywana liczba kontaktów dla pięcioaminokwasowego okna wyliczona według macierzy zdefiniowanej w (Miyazawa i Jernigan, 1999),
- informacja dotycząca obecności granic w sekwencjach homologicznych wykrytych przez program Hhsearch (biorąc pod uwagę przyrównanie do sekwencji homologicznych sprawdzane jest czy odpowiadająca reszta aminokwasowa znajduje się w pobliżu granicy domeny określonej w bazie CATH),
- macierz PSSM.

W tabeli 12 przedstawiono skuteczność przewidywania granic domenowych dla programu DomainSVM (obie warstwy) dla dwóch zbiorów testowych. Także w tym przypadku DomainSVM okazał się najlepszą metodą jednak różnica między DomainSVM, a FIEFdom nie jest tak duża jak w przypadku testów dla pierwszej warstwy.

Analizując wyniki prezentowane ze wspomnianej tabeli należy zwrócić uwagę na interesującą zależność, otóż wartości czułości są identyczne dla obu zbiorów testowych. Wynika to z faktu, że ilość wyników TP i FN nie zmienia się jeśli do zbioru singleANDmulti dodamy jedynie białka jednodomenowe (dla przypomnienia wynik TP oznacza poprawne wykrycie granicy, a więc poszerzając zbiór danych o białka nie posiadające reszt aminokwasowych należących do obszarów granic potencjalna pula reszt, które można zakwalifikować jako wyniki TP nie zmienia się; z analogiczną sytuacją mamy do czynienia rozpatrując pule wyników FN).

Tabela 12. Porównanie programu DomainSVM z innymi programami przewidującymi granicę domen przy uwzględnieniu reguły ± 20 aminokwasów. Skrótem „Spec" oznaczono specyficzność.

Metoda	**Zbiór danych**					
	singleANDmulti			**90_1000_CATH35**		
	Czułość	**Spec**	**ACC**	**Czułość**	**Spec**	**ACC**
DomainSVM	0,9545	0,8528	0,8195	0.9545	0,7364	0,7364
FIEFDom	0,9556	0,8449	0,8130	0,9556	0,7553	0,7297
DomPred (DomSSEA)	0,6667	0,6979	0,5174	0,6667	0,4810	0,3877
DomPred (DPS)	0,6127	0,5918	0,4307	0,6127	0,4125	0,3272
PPRODO	0,6288	0,4570	0,3599	0,6288	0,2825	0,2421
Scooby-domain (Domcut)	0,6925	0,2961	0,2617	0,6925	0,1818	0,1682
DoBo	0,5757	0,3156	0,2560	0,5757	0,2406	0,2072
Scooby-domain (Greedy)	0,5661	0,3108	0,2510	0,5661	0,2222	0,1899
DOMpro	0,1668	0,4407	0,1377	0,1668	0,3195	0,1231
GLOBPLOT	0,1148	0,3115	0,0915	0,1148	0,2214	0,0818

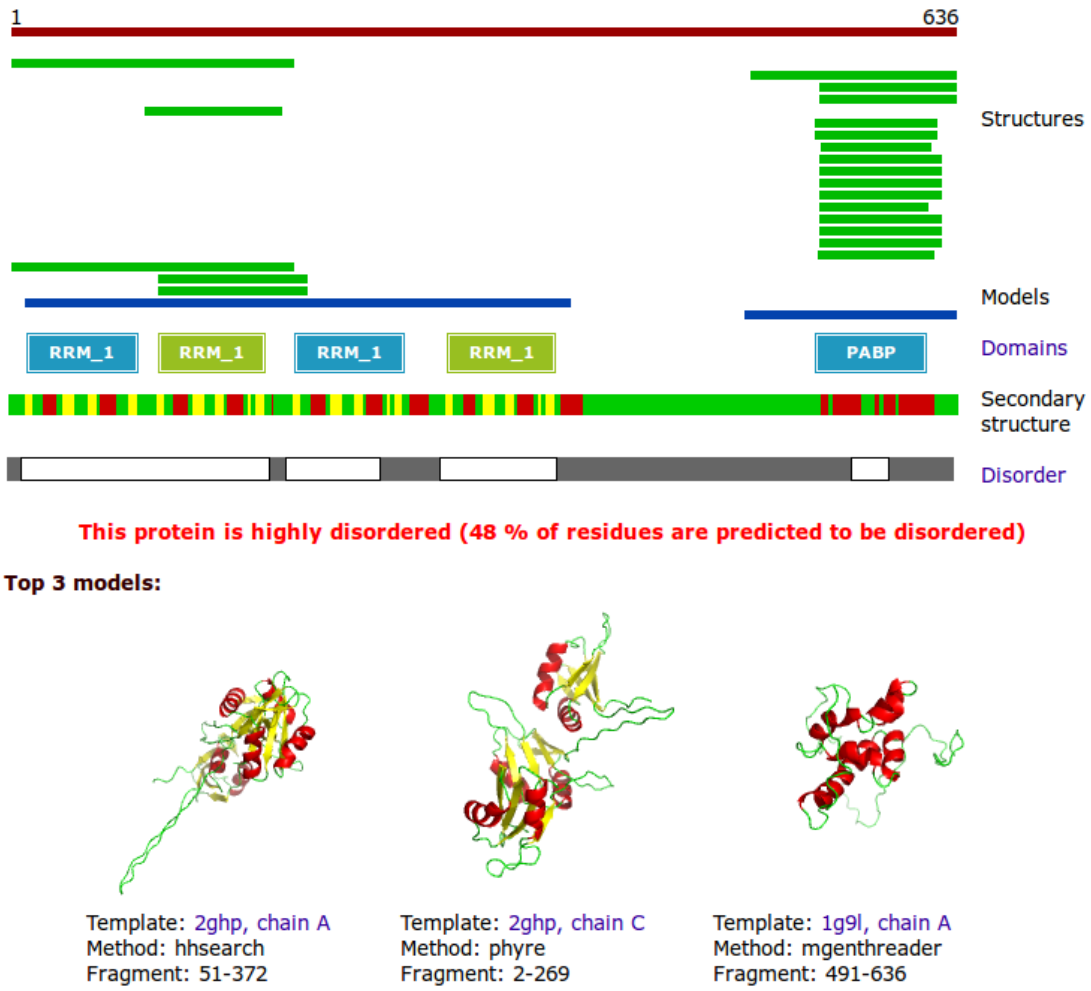
6.4. Analiza ludzkich białek odpowiedzialnych za modyfikację końca 3′ mRNA

Białka odpowiedzialne za cięcie i poliadenylację końca 3′ mRNA tworzą złożony kompleks zbudowany z około 30 białek stanowiących rdzeń kompleksu (rys. 9) oraz z ponad 50 białek pomocniczych. Dodatkowo liczba ta znacznie wzrasta jeśli uwzględnić izoformy poszczególnych białek powstałe w wyniku różnicowego składania transkryptów. Dla części z tych białek lub ich domen udało się rozwiązać strukturę metodą krystalografii rentgenowskiej (tabela 13), jednak ciągle wiedza na temat tych białek jest niewystarczająca. Należy zwrócić uwagę, że większość białek kompleksu można sklasyfikować jako białka duże, prawdopodobnie, zbudowane z kilku domen. Często rozwiązane struktury konkretnego białka, nawet jeśli jest ich dużo, reprezentują jedną lub więcej domen białka, rzadko kiedy pokrywają całą długość sekwencji białka (ryc. 20).

Korzystając z opisanych w niniejszej rozprawie programów przeprowadzono wnikliwą analizę białek odpowiedzialnych za modyfikację końca 3′ mRNA. W pierwszej kolejności wszystkie sekwencje ludzkiego kompleksu (uwzględniono także alternatywne izoformy, w sumie 60 białek) zostały wysłane na metaserwer GeneSilico. Ze względu na ograniczenie dotyczące długości siedem sekwencji przekraczających 990 reszt aminokwasowych podzielono na dwie części odpowiadające końcom N i C. Dostęp do poszczególnych wyników jest możliwy po zalogowaniu na metaserwer GeneSilico (login: mrna i hasło: mrna).

W czasie analizy skupiono się na przewidywaniu domen białkowych oraz wewnętrznego nieuporządkowania. Dodatkowo, opierając się na najlepszych szablonach wykrytych przez metaserwer dla każdego z białek wygenerowano modele homologiczne (wykorzystano w tym celu program MODELLER).

Niniejsza analiza ma za zadanie usystematyzować i poszerzyć wiedzę na temat ludzkich białek odpowiedzialnych za modyfikację końca 3′ mRNA. Wszystkie prezentowane wyniki zostały dodatkowo udostępnione w formie interaktywnej bazy danych pod adresem http://www.genesilico.pl/mrna3db/.



Ryc. 20. Informacja strukturalna dotycząca białka PABP-1 dostępna w internetowej bazie danych stworzonej w ramach niniejszej rozprawy. Idąc od góry, na zielono schematycznie zaznaczono położenie fragmentów odpowiadających znanym strukturom w bazie PDB. Na niebiesko zaznaczono fragmenty odpowiadające modelom zdeponowanym w bazie SWISS-MODEL Repository. Poniżej znajdują się informacja o domenach (według bazy Pfam). Następnie pokazana została struktura drugorzędowa (konsensus z przewidywań struktury drugorzędowej z metaserwera GeneSilico; kolorem czerwonym zaznaczono regiony helikalne, kolorem żółtym fragmenty odpowiadające wstęgom β, a kolorem zielonym pozostałe elementy struktury drugorzędowej). Poniżej znajduje się przewidywanie wewnętrznego nieuporządkowania według programu MetaDisorder (kolor szary regiony IUR, kolor biały regiony uporządkowane). Na samym dole przedstawiono trzy najlepsze modele wygenerowane programem MODELLER na podstawie szablonów zidentyfikowanych przez metaserwer GeneSilico.

Tabela 13. Wybrane statystyki dotyczące ludzkich białek odpowiedzialnych za modyfikację końca 3′ mRNA. iso – numer izoformy białka, mw – masa cząsteczkowa w kDa; ip – punkt izoelektryczny; % ss – procentowa zawartość elementów struktury drugorzędowej (łącznie helis α i wstęg β), konsensus z przewidywań struktury drugorzędowej z metaserwera GeneSilico; % IUR – procentowa zawartość regionów wewnętrznie nieuporządkowanych przewidzianych przez program MetaDisorderMD2; % lc – procentowa zawartość regionów o niskiej złożoności sekwencji wyznaczona za pomocą programu SEG; PDB – kody PDB znanych struktur (dla zwięzłości, w przypadku, gdy znane jest więcej niż dwie struktury podana są tylko dwa kody, a w nawiasie całkowita liczba struktur).

Nazwa	iso	długość	mw	ip	% ss	% IUR	% lc	PDB
CLP1	1	425	47,65	6,37	51,76	6,82	0,00	
	2	361	40,70	6,29	55,68	11,63	0,00	
CPSF1	1	1443	160,88	6,14	42,13	18,99	17,12	
CPSF2	1	782	88,49	4,85	50,13	28,64	11,25	
CPSF3	1	684	77,49	5,36	54,68	14,33	8,33	2I7T, 2I7V
CPSF4	1	269	30,25	8,12	8,55	45,35	26,77	2RHK, 2D9N
	2	244	27,55	8,03	9,43	38,11	0,00	
	3	243	27,42	8,03	10,29	41,98	0,00	
CPSF5	1	227	26,23	9,05	44,49	16,30	0,00	3MDI, 3BAP (11)
CPSF6	1	551	59,21	6,90	15,79	82,03	67,33	3Q2S, 3Q2T (5)
	2	588	63,47	7,56	12,24	85,37	63,27	
CPSF6	3	478	52,33	6,18	16,11	80,33	62,13	
CPSF7	1	471	52,05	7,95	17,20	78,77	45,22	3N9U
	2	462	51,10	7,66	18,18	78,14	44,59	
	3	514	56,37	9,05	19,07	81,32	48,44	
CSTF1	1	431	48,36	6,30	46,64	6,96	0,00	
CSTF2	1	577	60,96	6,56	23,57	79,20	61,70	2J8P, 1P1T
	2	560	59,25	6,50	24,11	77,68	60,89	
CSTF2T	1	616	64,44	7,00	22,08	79,38	66,07	
CSTF3	1	717	82,92	8,06	64,44	21,90	23,99	
	2	103	12,10	6,94	75,73	18,45	0,00	
FIP1L1	1	594	66,53	5,35	9,60	90,40	77,61	
	3	520	58,38	5,34	11,92	87,31	73,65	
	4	378	40,83	4,43	10,32	85,19	67,72	
PABP-1	1	636	70,67	9,81	43,08	48,27	27,67	1CVJ, 4F02 (20)
	2	547	61,18	9,35	47,53	40,22	21,39	
PABP-2	1	306	32,75	4,89	29,08	73,20	75,82	3B4D, 3B4M (3)
	2	296	31,50	4,78	29,73	72,97	82,77	
	3	333	37,17	8,08	53,15	39,94	17,72	
PABP-3	1	631	70,03	10,0	44,06	47,54	23,93	2D9P
PABP-4	1	644	70,78	9,56	44,72	46,27	33,54	
	2	631	69,58	9,86	43,90	45,32	32,17	
	3	660	72,39	9,66	41,52	50,15	37,73	
PABP-5	1	382	43,33	9,84	53,93	17,02	0,00	
PAPD1	1	582	66,17	9,19	41,75	33,51	9,28	3PQ1
	2	712	78,83	8,14	36,94	45,65	32,87	
PAPD4	1	484	56,03	9,65	37,60	31,61	22,52	
	2	480	55,50	9,72	37,71	30,42	13,75	
PAPOLA	1	745	82,84	7,14	35,70	40,13	40,81	
	2	285	32,63	7,61	56,14	9,12	24,91	
PAPOLB	1	636	71,68	6,18	42,77	27,52	12,11	
PAPOLG	1	736	82,80	9,47	36,14	36,55	20,38	
PCF11	1	1555	173,05	8,52	14,02	86,82	69,20	
PP-1A	1	330	37,51	6,05	46,67	10,91	0,00	3N5U, 4G9J (8)
	2	341	38,63	6,35	44,87	14,96	0,00	
PP-1B	1	327	37,19	5,91	47,71	10,09	0,00	

Nazwa	iso	długość	mw	ip	% ss	% IUR	% lc	PDB
RBBP6	1	1792	201,56	10,03	6,25	91,57	76,34	2C7H, 2YSA (4)
	2	1758	197,29	10,01	6,71	91,30	76,00	
	3	118	13,24	9,02	40,68	24,58	0,00	
	4	952	106,04	9,38	9,56	84,03	72,48	
RPB1	1	1970	217,18	7,10	36,45	30,10	22,28	2GHQ, 2GHT
SSU72	1	194	22,57	5,08	59,28	2,06	0,00	3O2Q, 3O2S (4)
	2	153	16,93	9,70	41,18	8,50	0,00	
Symplekina	1	1274	141,15	5,88	54,55	33,91	21,43	3ODR, 3ODS (7)
	2	673	74,52	7,30	55,13	38,34	14,56	
TUT1	1	874	93,85	5,90	36,61	59,38	69,57	2E5G
WDR33	1	1336	145,89	9,49	12,50	72,01	58,68	
	2	326	38,29	9,77	42,02	11,96	0,00	
	3	257	30,29	9,83	41,63	12,45	0,00	
WDR82	1	313	35,08	7,56	50,16	0,32	0,00	
Średnia	1,65	608	67,68	7,66	32,09	51,43	38,16	

* w kolumnach oznaczonych %ss, %IUR oraz %lc podano średnią globalną liczoną na poziomie reszt aminokwasowych

6.4.1. Przewidywanie regionów wewnętrznie nieuporządkowanych białek odpowiedzialnych za modyfikację końca 3′ mRNA

Jedną z charakterystycznych cech białek wewnętrznie nieuporządkowanych jest tendencja do występowania w wieloskładnikowych kompleksach makrocząsteczek (Haynes et al., 2006a; Kim et al., 2008). Tłumaczy się to tym, że zwiększona labilność takich białek umożliwia dostosowanie się regionów IUR do innych składników kompleksu, które zmieniają się w czasie i przestrzeni. W ten sposób regiony wewnętrznie nieuporządkowane dostosowują swój kształt do aktualnych potrzeb. Dzięki takiej możliwości mogą one oddziaływać z większą liczbą białek lub/i pełnić bardziej złożone funkcje (Dunker et al., 2005).

W ramach przeprowadzonej w niniejszym projekcie badawczym analizy ludzkich białek odpowiedzialnych za modyfikację końca 3′ mRNA szczególny nacisk położono na przewidywanie występowania regionów wewnętrznie nieuporządkowanych. Zasadniczym problemem w przypadku białek omawianego kompleksu jest niepewność co do dokładnego składu kompleksu. Na podstawie przeprowadzonych badań wiadomo jakie białka stanowią rdzeń kompleksu, jednak nie zawsze wiadomo na jakim etapie procesu dane białka są aktywne oraz czy część z nich nie jest zastępowana przez inne. Przykładowo, w literaturze praktycznie brak jest informacji na temat specyficzności poszczególnych izoform powstałych w wyniku różnicowego składania transkryptów. Ponadto, przynajmniej część białek kompleksu działa w formie dimerów (np. białka CPSF5, CSTF2).

Aby zniwelować tą niepewność białka odpowiedzialne za modyfikację końca 3′ mRNA podzielono na trzy grupy. W skład pierwszej z nich (tzw. kompleks A) wchodzą wszystkie białka wliczając w to alternatywne izoformy (w sumie 60 białek). W skład drugiej grupy (tzw. kompleks B) wchodzą jedynie poszczególne białka (w przypadku, gdy białko posiada więcej niż jedną izoformę, zawsze wybierane jest białko oznaczone jako izoforma pierwsza, najczęściej występująca w komórce). Dodatkowo stworzono trzecią grupę (tzw. kompleks C) której skład odzwierciedla aktualny stan wiedzy na temat stechiometrii kompleksu. W jej skład wchodzą następujące białka: symplekina, PAPOLA, PP-1A, PABP-1, WDR82, SSU72, RBBP6, CPSF1, CPSF3, CPSF2, CPSF4, FIP1L1, WDR33, CSTF2, CSTF2T, CSTF3, CSTF1, CSTF3, CSTF1, CPSF6, CPSF7, CPSF5, CPSF5, PCF11, CLP1 oraz koniec C białka RPB1 (od reszty aminokwasowej 1478 do końca białka). Należy zwrócić uwagę, że z jednej strony w przypadku białek tworzących dimery odpowiednie białka występują więcej niż jeden raz, natomiast z drugiej strony część białek blisko spokrewnionych (np. białka PABP-1 do PABP-5) reprezentowana jest jedynie przez jedno białko. Specjalnym przypadkiem jest białko RPB1, czyli polimeraza RNA typu II. Jest to duże białko zbudowane z blisko 2000 reszt aminokwasowych. Jego podstawową funkcją, za którą odpowiadają ściśle ustrukturalizowane domeny (N-końcowych 1500 reszt aminokwasowych), jest synteza RNA. Dodatkowo, białko to posiada C-końcową domenę, która oddziałuje z licznymi białkami kompleksu białek odpowiedzialnych za modyfikację końca 3′ mRNA. Ogon ten zawiera 52 heptapeptydowe powtórzenia o sekwencji konsensusowej YSPTSPS i jest wewnętrznie nieuporządkowany. Mając na uwadze specyfikę białka RPB1 w skład kompleksu zaliczono jedynie koniec C. Dodatkowo, wiadomo, że domena C-końcowa podlega intensywnej modyfikacji posttranslacyjnej (Egloff i Murphy, 2008) za co odpowiadają białka PP-1A oraz SSU72 (również wliczone w poczet kompleksu C).

W tabeli 14 przedstawiono wyniki dotyczące występowania regionów wewnętrznie nieuporządkowanych. Bez względu jak zdefiniujemy analizowany kompleks, regiony wewnętrznie nieuporządkowane występują ponad dwukrotnie częściej niż w innych białkach proteomu ludzkiego (średnio stanowią one około połowy ich długości). Ponadto ponad 80% białek proteomu ludzkiego posiada przynajmniej jeden region wewnętrznie nieuporządkowany dłuższy niż 30 reszt aminokwasowych. Otrzymany wynik potwierdza ostatnie badania, według których białka oddziałujące z mRNA, na wszystkich etapach jego trwania, cechują się wysokim stopniem nieuporządkowania (Castello et al., 2012). Ponadto przewidywana liczba

wewnętrznego nieuporządkowania jest bardzo podobna do opisanej dla białek ludzkiego spliceosomu (Korneta i Bujnicki, 2012; Korneta et al., 2012), innego wielobiałkowego, dynamicznego kompleksu oddziałującego z mRNA.

Tabela 14. Statystyki dotyczące występowania regionów wewnętrznie nieuporządkowanych dla białek odpowiedzialnych za modyfikację końca 3′ mRNA (kompleks A oznacza wszystkie białka uwzględniając izoformy; kompleks B oznacza białka reprezentujące pierwszą izoformę; kompleks C uwzględnia dane na temat stechiometrii kompleksu – szczegóły w tekście). Dodatkowo dla porównania dane te zestawiono z analogicznymi danymi dotyczącymi białek spliceosomu ludzkiego oraz całego proteomu człowieka. Oznaczenia: „% IUR" – procentowa zawartość regionów wewnętrznie nieuporządkowanych przewidzianych przez program MetaDisorderMD2; „IUR > 30" – procentowa ilość białek zawierających przynajmniej jeden region wewnętrznie nieuporządkowanych o długości powyżej 30 reszt aminokwasowych; „IUR > 50" – procentowa ilość białek zawierających przynajmniej jeden region wewnętrznie nieuporządkowanych o długości powyżej 50 reszt.

Grupa białek	Liczba białek	% IUR	IUR > 30	IUR > 50
kompleks A	60	51,43	81,67 (49 białek)	73,33 (44 białek)
kompleks B	33	47,30	81,82 (27 białek)	75,76 (25 białek)
kompleks C	26	50,44	80,77 (21 białek)	80,77 (21 białek)
spliceosom[1]	122	53,5 (44,0)	80,30 (98 białek)	NA
proteom ludzki[2]	26385	21,6	35,20 (~9300 białek)	21,9 (~5800 białek)

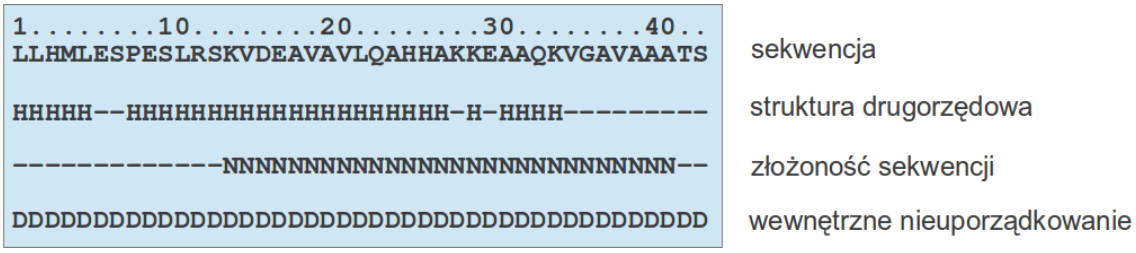
[1] na podstawie (Ward et al., 2004)
[2] na podstawie (Korneta i Bujnicki, 2012)

Ogólnie, regiony wewnętrznie nieuporządkowane stanowią dość niejednorodną grupę. W ich obrębie możemy znaleźć takie, które cechują względnie łatwym przejściem ze stanu braku uporządkowania w stan ustrukturalizowany oraz takie, które raczej zawsze pozostają wewnętrznie nieuporządkowane (Vucetic et al., 2003). Wykorzystanie przewidywań struktury drugorzędowej i przewidywanie regionów o niskiej złożoności może pomóc nam odróżnić od siebie te dwie klasy wewnętrznego nieuporządkowania. Tabela 15 zawiera przykładowe dane dla regionów wewnętrznie nieuporządkowanych, które cechują się skrajnymi wartościami tych parametrów. W obrębie 60 analizowanych białek (wliczając izoformy) można zidentyfikować w sumie 253 oddzielnych fragmentów przewidywanych jako wewnętrznie nieuporządkowane i których długość jest większa niż 20 reszt aminokwasowych. Najdłuższy z nich, należący do białka RBBP6, ma długość 1470 reszt aminokwasowych (92% długości całego białka przewidywanych jest jako wewnętrznie nieuporządkowane). Na podstawie otrzymanych wyników można zaobserwować ciekawą prawidłowość. O ile względnie łatwo można znaleźć regiony nieuporządkowane cechujące się silną tendencją do tworzenia struktury drugorzędowej (potencjalne regiony IUR, które mogą łatwo ulegać ustrukturalizowaniu np. region 602-645 w białku PABP-4) to sytuacja jest bardziej skomplikowana, jeśli dodatkowo w analizie

uwzględnimy regiony o niskiej złożoności sekwencji. Po pierwsze jeśli region wewnętrznie nieuporządkowany zawiera wysoki odsetek struktury drugorzędowej, to prawie zawsze klasyfikowany jest jako sekwencja o wysokiej złożoności. Z kolei, jeśli dany region wewnętrznie nieuporządkowany klasyfikowany jest jako sekwencja o niskiej złożoności to ilość potencjalnych elementów struktury drugorzędowej jest znikoma, a nawet jeśli jakieś występują to są to prawie zawsze helisy α. Sytuację tą zobrazowano na Ryc. 21.

Tabela 15. Przykładowe dane dla fragmentów białek, które zostały przewidziane jako wewnętrznie nieuporządkowane i jednocześnie cechują się skrajnymi wartościami przewidywanej struktury drugorzędowej oraz złożoności sekwencji. Podano po trzy przykłady fragmentów białek o kolejno największej długości, z najwyższą zawartością struktury drugorzędowej oraz o największej procentowej zawartości regionów o niskiej złożoności. Oznaczenia jak w tabeli 13. Pełna wersja prezentowanej tabeli zawiera ponad 250 wierszy i została udostępniona w formie interaktywnej (z funkcją sortowania danych w kolumnach) na stronie http://www.genesilico.pl/mrna3db/3mrna_complex_disorder.html.

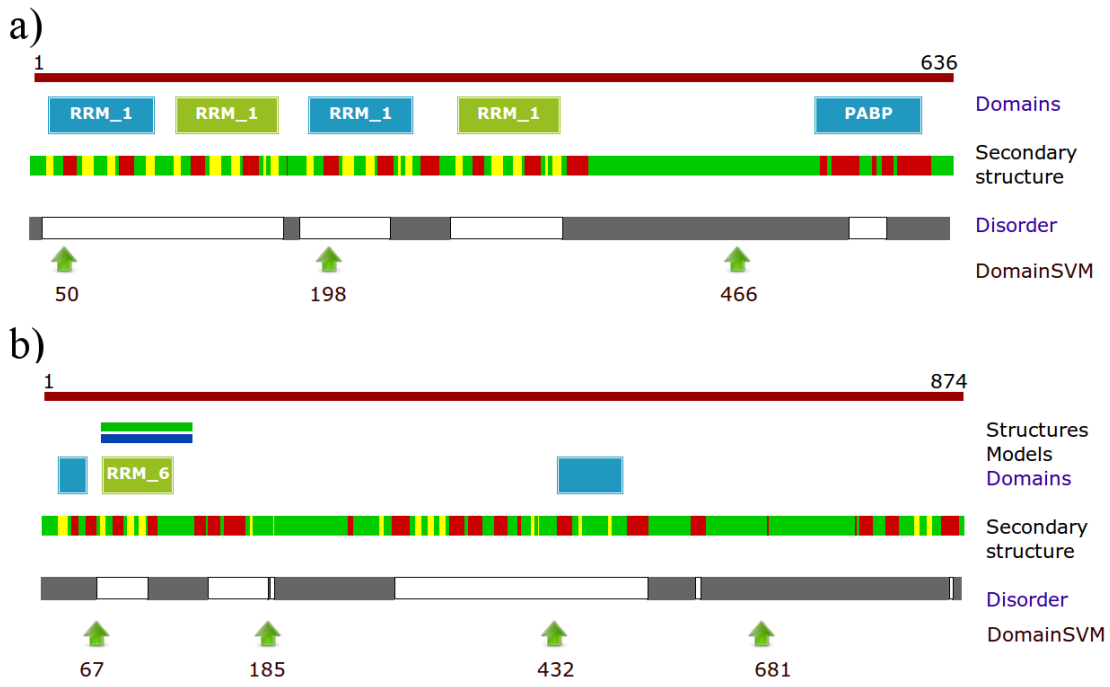
Nazwa	Lokalizacja	Długość	% ss	% helis α	% wstęg β	% lc
RBBP6	323-1793	1470	1,77	1,77	0,00	87,96
WDR33	418-1336	918	1,96	1,53	0,44	85,29
PCF11	693-1324	631	1,90	1,90	0,00	86,21
PABP-4	269-291	22	86,36	86,36	0,00	0,00
PABP-4	602-645	43	69,77	69,77	0,00	65,12
CPSF3	632-685	53	62,26	33,96	28,30	0,00
TUT1	102-159	57	35,09	35,09	0,00	100,00
CPSF6	459-552	93	27,96	27,96	0,00	100,00
Symplekina	330-400	70	24,29	24,29	0,00	100,00



Ryc. 21. Przewidywania struktury drugorzędowej, złożoności sekwencji oraz wewnętrznego nieuporządkowania dla regionu 602-645 białka PABP-4. Region ten cechuje się nadzwyczajnie dużą zawartością struktury drugorzędowej przy jednocześnie niskiej złożoności sekwencji. „H' – helisa α, „N" – reszty aminokwasowe o niskiej złożoności sekwencji, „D" – reszty aminokwasowe przewidziane jako wewnętrznie nieuporządkowane.

6.4.2. Domeny białek odpowiedzialnych za modyfikację końca 3′ mRNA

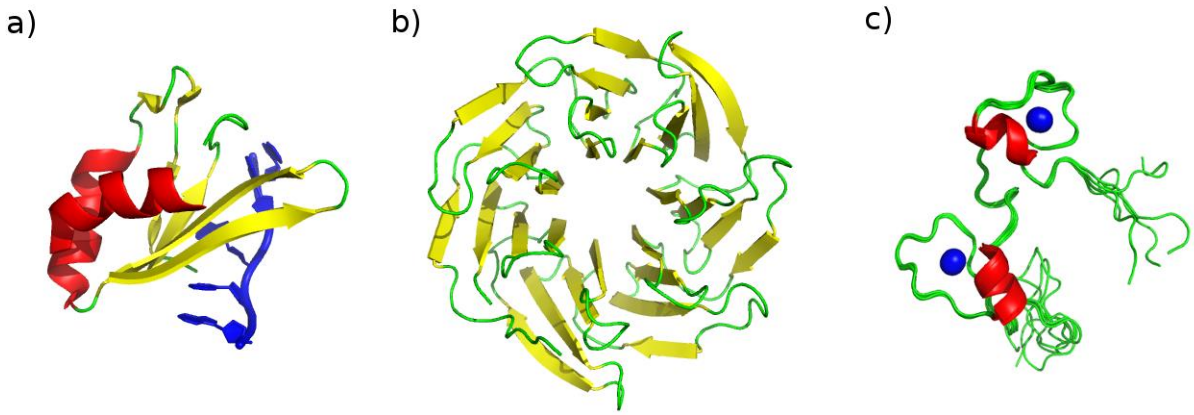
Przeprowadzona w ramach niniejszego projektu analiza domen białek odpowiedzialnych za modyfikację końca 3′ mRNA polegała na pobraniu z bazy PFAM informacji dotyczącej znanych domen i porównaniu jej do obszarów przewidywanych jako wewnętrznie nieuporządkowanych oraz granic domen przewidzianych za pomocą programu DomainSVM. Przykładowy wynik dla białek PABP-1 i TUT1 przedstawiono na ryc. 22. Jak widać granice domen, które przewidziane są przez program DomainSVM zwykle zlokalizowane są w okolicy jednej z granic domen zdefiniowanych w bazie PFAM (np. pozycje 50 i 198 dla biała PABP-1 czy też pozycje 67 i 432 dla białka TUT1). Innym częstym przypadkiem jest przewidywanie granicy domen w miejscach w których część białka ustrukturalizowanego przechodzi w region wewnętrznie nieuporządkowany (pozycja 185 w białku TUT1). Oczywiście, nie zawsze tak jest (pozycja 466 w białku PABP-1). Z jednej strony może to być błąd programu (wynik fałszywie dodatni, błąd I rodzaju), ale może to też być istotna informacja służącą jako punk startowy do dalszych badań.



Ryc. 22. Znane domeny, przewidywane granice domen oraz regiony wewnętrznie nieuporządkowane na przykładzie białek PABP-1 (a) oraz TUT1 (b). Zielonymi strzałkami zaznaczono miejsca w których program DomainSVM przewiduje granicę między domenami. Kolory i oznaczenia jak na ryc. 20.

W czasie prezentowanych badań zanalizowano skład domen, które występują w białkach odpowiedzialnych za modyfikację końca 3′ mRNA. Najczęściej występujące grupy domen białkowych to:

- domeny RRM (ang. *RNA recogniction motif*) – sumarycznie 22 domeny w białkach TUT1, CPSF6 oraz PABP-1 do PABP-5; domeny te zwykle zbudowane są z koło 90 reszt aminokwasowych i posiadają zdolność do wiązania się z jednoniciowym RNA,
- domeny WD40 – sumarycznie 15 domen w białkach WDR33, WDR82 i CSTF1; domeny te zwykle zbudowane są w powtarzalnych motywów WD40 zakończonych resztami tryptofanu (W) i kwasu asparaginowego (D), ich średnia długości wynosi 40 reszt aminokwasowych, tworzą one charakterystyczną strukturę nazywaną śmigłem β (ang. *beta-propeller*), której funkcja związana z tworzeniem oddziaływań między białkami,
- palce cynkowe (ang. *zinc finger domain*) – sumarycznie 6 domen w białkach TUT1, RBBP-6 i CPSF4; domena ta jest stosunkowo krótka i jest zbudowana z dwóch antyrównoległych wstęg β i helisy α, do jej działania niezbędna jest obecność jonu cynku (Zn^{2+}); domeny te występują powszechnie u białek oddziałujących z DNA i RNA, w analizowanych białkach najczęściej występowały palce cynkowe typu C-x8-C-x5-C-x3-H (ryc. 23).



Ryc. 23. Trzy typy najczęściej występujących domen w białkach odpowiedzialnych za modyfikację końca 3′ mRNA: a) domena RRM białka PABP-1 (fragment 1-190) oddziałująca z mRNA – kolor niebieski (kod pdb: 4F02); b) model domeny śmigła β zbudowanego z domen WD40 białka WDR33 (fragment 1-405), którego podstawą był szablon z białka o kodzie pdb 1VYH; c) palec cynkowy białka CPSF4 (fragment 61-126), pokazano pięć nałożonych na siebie modeli otrzymanych techniką NMR, kolorem niebieskim zaznaczono jony cynku (kod: 2D9N).

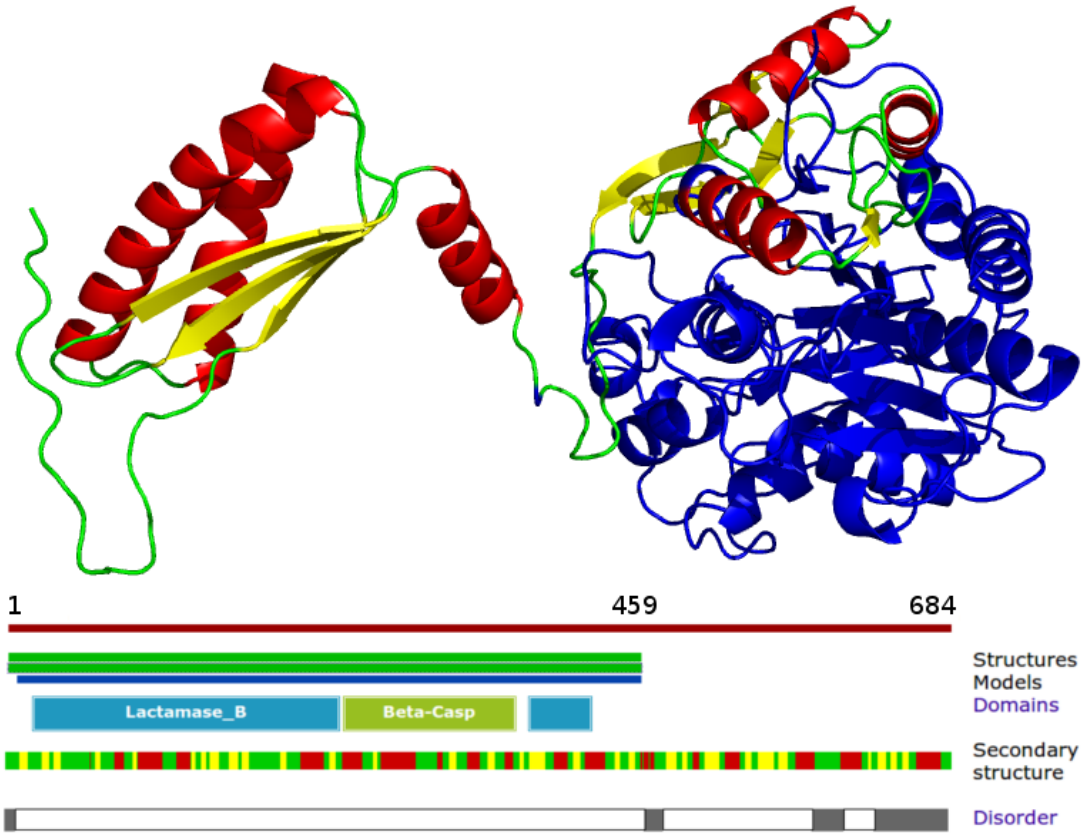
Ilościowa analiza wszystkich białek kompleksu pozwala wysunąć następujące wnioski (tabela 16):

- regiony odpowiadające domenom opisanym w bazie PFAM prawie zawsze zlokalizowane są w regionach, które zostały przewidziane przez program MetaDisorderMD2 jako uporządkowane,
- regiony przewidziane jako wewnętrznie nieuporządkowane najczęściej występują poza obszarami domen,
- regiony odpowiadające domenom, w przeciwieństwie do regionów leżących poza nimi, cechują się wysoką złożonością sekwencji,
- zawartość elementów struktury drugorzędowej miedzy obszarami domen a obszarami leżącymi poza nimi jest na zbliżonym poziomie.

Struktury przestrzenne wszystkich białek analizowanego kompleksu zostały wymodelowane za pomocą programu MODELLER w oparciu o szablony zidentyfikowane przez metaserwer GeneSilico. Wśród nich szczególnie interesujące mogą być modele białek, w obrębie których można zlokalizować potencjalnie nieustrukturalizowane regiony (według przewidywań programu MetaDisorder), a dla których struktura nie została jeszcze rozwiązana. Przykładem może być białko CPSF3 (ryc. 24).

Tabela 16. Ogólne statystyki domen wchodzących w skład kompleksu odpowiedzialnego za modyfikację końca 3′ mRNA. Poprzez obszary domenowe należy rozumieć regiony białka odpowiadające domenom oznaczonym w bazie PFAM, pozostałe regiony wliczone zostały na poczet obszarów pozadomenowych. Pod uwagę wzięte zostały jedynie izoformy główne. Oznaczenia jak w tabeli 13.

	% całości	**% IUR**	**% ss**	**% helis α**	**% wstęg β**	**% lc**
Obszary domenowe	43,78	4,49	23,26	14,35	8,91	3,72
Obszary pozadomenowe	56,22	72,68	21,25	15,29	5,96	55,18



Ryc. 24. Model homologiczny białka CPSF3. Na niebiesko zaznaczono N-końcowych 459 reszt aminokwasowych pochodzących ze struktury rozwiązanej technikami doświadczalnymi (kod pdb: 2I7T). Brakujący fragment między resztami 460-684 wymodelowano w oparciu o szablon pochodzący z białka o kodzie 3ZQ4. Dobudowany fragment możemy podzielić na dwie części. Pierwsza z nich „opłaszcza" domenę endonukleazy (przewidywanie według programu MetaDisorderMD2 klasyfikuje ten fragment jako ustrukturalizowany). Dalej znajduje się krótki, na wpół helikalny odcinek łączący zgrupowanie dwóch helis α i trzech wstęg β (fragment ten został przewidziany jako wewnętrznie nieuporządkowany, jednak przewidywanie struktury drugorzędowej sugeruje, że w regionie tym mogą być zlokalizowane elementy struktury drugorzędowej, które zostały automatycznie dobudowane przez program MODELLER, prawdopodobnie oznacza to, że fragment ten jest przykładem regionu wewnętrznie nieuporządkowanego, który stosunkowo łatwo przechodzi w stan ustrukturalizowania, przyjmując specyficzny kształt). Ponadto, interesujące jest to, że fragment ten przypomina domenę RMM (ryc. 23a), kształt i topologia są zbliżone brak jest jedynie jednej wstęgi β.

7. Dyskusja

W niniejszej rozprawie zaprezentowano metaserwer GeneSilico będący wygodną w użyciu platformą za pomocą której użytkownik może uruchomić niespotykaną nigdzie indziej liczbę programów bioinformatycznych. Serwis ten, dzięki ujednoliceniu formatu oraz intuicyjnemu interfejsowi, pozwala na łatwą interpretację wyników i nie wymaga od użytkownika wiedzy z dziedziny bioinformatyki.

Ponadto przedstawiono wyniki prac nad narzędziami do przewidywania wewnętrznego nieuporządkowania oraz domen w białkach, których efektem końcowym były dwa zupełnie nowe programy. Programy te intensywnie przebadano pod kątem statystycznej istotności generowanych wyników, a ich bezpośrednim testem była przeprowadzona przez autora rozprawy analiza białek odpowiedzialnych za modyfikację końca 3′ mRNA.

W dalszej części dyskusji przedstawione zostanie krytyczne omówienie otrzymanych wyników, użytej metodologii oraz zaprezentowane zostaną przyszłe kierunki badań.

7.1. Przewidywania regionów wewnętrznie nieuporządkowanych

Przewidywanie regionów wewnętrznie nieuporządkowanych jest istotnym problemem badawczym zarówno ze względu na ich znaczenie biologiczne (białka posiadające regiony IUR związane są z procesami takimi jak: proliferacja, apoptoza, powstawanie nowotworów, przekazywanie sygnału oraz regulacja transkrypcji i translacji) jak i „techniczne” (często rozwiązanie struktury białka metodą krystalografii rentgenowskiej nie jest możliwe, ponieważ nie można otrzymać kryształów wystarczającej jakości, za co mogą odpowiadać regiony wewnętrznie nieuporządkowane obecne w docelowym białku).

W pierwszym etapie badań skonstruowano meta-metodę, która wykorzystuje następujące programy: DisEMBL, GLOBPLOT, DISOPRED2, DISPROT (VSL2B), IUPred (uruchamiany w dwóch trybach dostosowanych do przewidywania długich i krótkich regionów wewnętrznie nieuporządkowanych), DISpro, RONN, SPRITZ, PDISORDER, POODLE-L, POODLE-S, PrDOS oraz iPDA. Jako wynik końcowy zwracany jest konsensus ważony (każda metoda składowa otrzymuje swoją wagę, tym większą im lepsze program zwraca wyniki). Wagi przypisane programom zoptymalizowano opierając się na specjalnie przygotowanym zbiorze testowym oraz używając sztucznych sieci neuronowych. Wyniki zaprezentowane w tabeli 4

jednoznacznie pokazują, że otrzymana meta-metoda była lepsza niż dowolny jej element składowy (miara AUC wynosiła odpowiednio 0,868 dla programu FloatCons i 0,83 dla programu DISPROT). Wersja programu oparta na algorytmie ANN była najlepsza w przypadku miary S_w (choć należy podkreślić, że miary tej użyto w procesie optymalizacji). Zgodnie z oczekiwaniem program BinCons uwzględniający jedynie wartość binarną ('D" lub „O") był nieznacznie gorszy niż FloatCons, który otrzymywał pełną informację (prawdopodobieństwo bycia w stanie nieuporządkowania dla danej reszty aminokwasowej). Wynik ten dotyczył dwóch z trzech miar (ocena według miary MCC była najwyższa).

Z powyższymi wynikami trzeba skonfrontować wyniki niezależnego testu jakim był eksperyment CASP8 (tabela 6). Brało w nim udział w sumie 25 programów, wśród których BinCons i FloatCons (oznaczone odpowiednio numerami 153 i 297) okazały się najlepszymi metodami bez względu na kryterium oceny. Dość zaskakujący może być fakt, że według tego testu różnica między BinCons i FloatCons była nieistotna statystycznie (w zależności od miary oceny lepsza była raz jedna raz druga wersja). Ponadto, wynik ANNCons jest wyraźnie słabszy niż wynik otrzymany w czasie wcześniejszych testów. Prawdopodobnie odzwierciedla to stan przeuczenia metody opartej na sieci neuronowej względem wartości S_w i konkretnego zbioru testowego.

W następnym etapie badań postanowiono wzbogacić metodę FloatCons o informację dotyczącą występowania przerw w szablonach wykrytych przez programy do rozpoznawania zwoju. W efekcie stworzono następujące warianty programu:

- MetaDisorder3D – oparty wyłącznie na występowaniu przerw w szablonach, zasadniczym problemem w jego przypadku było uwzględnienie wpływu jakości przyrównania szablonu oraz rodzaju użytego programu na jakość przewidywania. W tym przypadku użyto algorytmu genetycznego, który miał za zadanie znaleźć optymalne wagi dla poszczególnych programów dzieląc ich wyniki na trzy kategorie (przyrównania dobrej, średniej i wątpliwej jakości);
- MetaDisorderMD – oparty na połączeniu MetaDisorder3D z FloatCons, w tym przypadku problem polegał na optymalnym połączeniu obu komponentów, także w tym przypadku użyto algorytmu genetycznego;

– MetaDisorderMD2 – wariant ten zbudowano identycznie jak poprzedni, z tą różnicą, że do jego optymalizacji użyto specjalnie zaprojektowanej miary S_{ww}, która miała lepiej oceniać jakość przewidywań.

Jako zbioru treningowego użyto danych z eksperymentu CASP8 dla których skuteczność MetaDisorderMD i MetaDisorderMD2 przekraczała 90% (tabela 7). Za bardziej miarodajny test należy uznać wyniki eksperymentu CASP9, gdzie, choć metoda osiągnęła słabszy wynik (AUC=0.818, metoda PrDOS2 uzyskała AUC=0,855), to ciągle pod kątem miary S_w najlepsza (0,516 kontra 0,509). Należy zwrócić uwagę, że nie korzystająca z informacji pochodzącej z szablonów metoda FloatCons była wyraźnie słabsza, co ogólnie zgadza się z wynikami testów przeprowadzonych przez autora rozprawy.

Na cały proces powstawania różnych wersji programu do przewidywania regionów wewnętrznie nieuporządkowanych można patrzeć jak na serię eksperymentów, w których każdy kolejny został zaprojektowany w oparciu o wyniki poprzedniego lub jako jego kontrola. Przykładowo program MetaDisorderMD stanowiący połączenie wersji FloatCons z wersją MetaDisorder3D można porównać bezpośrednio do programów, z których został zbudowany. Już w momencie projektowania MetaDisorder3D widomo było, że będzie to klasyfikator gorszy niż istniejące programy do przewidywania wewnętrznego nieuporządkowania. Mając to na uwadze, można było ograniczyć się jedynie do testowania finalnej wersji MetaDisorderMD. Autor wiedziony ciekawością naukową chciał jednak uzyskać odpowiedz nie tylko na pytanie o ile poprawia się skuteczność nowej metody, ale także jak duży wpływ mają na siebie poszczególne jej elementy składowe. Podobnie postąpiono w przypadku wariantu MetaDisorderMD2, który miał za zadanie zbadać wpływ nowej miary oceny skuteczności przewidywania (miara S_{ww} jest jednym z wyników niniejszej rozprawy). Jako referencji można było tu użyć metody MetaDisorderMD.

Podsumowując tą część badań należy stwierdzić, że kolejne wersje metody MetaDisorder były najlepszymi programami do przewidywania wewnętrznego nieuporządkowania w eksperymencie CASP8 i CASP9, tak więc skuteczność programu została potwierdzona nie tylko przez testy autora rozprawy, ale także w czasie niezależnej oceny i w konfrontacji z ponad 20 innymi programami tego typu tworzonymi przez badaczy z całego świata.

7.2. Przewidywanie domen w białkach

Kolejnym wynikiem badań przeprowadzonych w ramach prezentowanej pracy doktorskiej jest program przewidujący występowanie domen. Celem badawczym było tutaj przewidywanie granic domen, ponieważ to one tak naprawdę określają domenę. Badacz poszukujący odpowiedzi na pytanie dotyczące budowy domenowej białka tak naprawdę zwykle ma na myśli nie określenie czy dana reszta aminokwasowa przynależy do konkretnej domeny, lecz informację odkąd dokąd ciągnie się sekwencja danej domeny. Oczywiście definicja domeny jako ciągłego elementu sekwencji nie uwzględnia dwóch jej ważnych aspektów. Po pierwsze domena odnosi się do regionu białka zdefiniowanego jako kształt trójwymiarowy. Oznacza to, że istnieją przypadki, gdy reszty aminokwasowe jednej domeny zlokalizowane są w innych regionach sekwencji. Choć więc taka „płaska" definicja domeny wydaje się nieodpowiednia, to zdecydowana większość domen (80%) ją spełnia (Jones et al., 1998). Po drugie, znane są także takie przypadki, gdy domena zbudowana jest z fragmentów pochodzących z dwóch niezależnych łańcuchów tego samego lub innego białka (ang. *domain swapping*). Znanych jest jednak zaledwie kilkaset białek posiadających takie domeny (Shameer et al., 2011b). Z tego względu dla uproszczenia problemu w programie DomainSVM rozpatrywane są jedynie domeny ciągłe.

Program DomainSVM opiera swoje działanie o maszynę wektorów nośnych (ang. *suport vector machine*). Aby ulepszyć skuteczność programu, analizowany problem podzielono na dwa mniejsze. Pierwsza warstwa programu ma za zadanie odróżnienie białek jednodomenowych od wielodomenowych, ponieważ zdecydowana większość białek w zbiorze testowym zbudowana była z białek jednodomenowych, a więc nie niosących informacji o granicach (koniec białka traktowany jest jako przypadek trywialny i nie jest rozpatrywany) – ryc. 19 i tabela 9.

W przypadku tej warstwy jako cech wejściowych użyto następujących cech: długość sekwencji, entropia Shannona i entropia metryczna, średnia hydrofobowość GRAVY z wykorzystaniem skali hydrofobowości Kyte-Doolittle'a (Kyte i Doolittle, 1982), przewidywanie wewnętrznego nieuporządkowania według programu RONN wyrażone w formie procentowej zawartości regionów IUR, obecność regionu IUR o długości przynajmniej 10, 20, 30, 40, 50 i 60 reszt aminokwasowych, liczba regionów IUR o długości powyżej 10 reszt, struktura drugorzędowa przewidziana za pomocą programów PSIPRED i Sspro4, relatywna dostępność dla rozpuszczalnika (RSA) przewidziana za pomocą programu ACCpro dla trzech progów 0%, 5% i 25% dostępności, informacja dotycząca liczby domen u najbliższych homologów (w

pierwszej kolejności za pomocą programu HHsearch identyfikowane jest 10 najlepszych sekwencji homologicznych, a następnie dla tych sekwencji pobierana jest informacja na temat liczby domen według bazy CATH). Największy wpływ na skuteczność przewidywania miała informacja dotycząca domen bliskich homologów (dokładność przewidywania na poziomie 95%). Jednak nawet, gdy brak jest takich danych, skuteczność programu pozostaje na akceptowanym poziomie wynoszącym 78% (tabela 10). Ciekawym wynikiem tej części badań jest względnie wysoki wpływ entropii. Inna wysoko oceniona cecha dotyczy korelacji między ilością domen a długością i jest powszechnie znana (Wheelan et al., 2000)). Wynik działania programu dla problemu dyskryminacji białek jednodomenowych i wielodomenowych porównano z 10 innymi programami. Jedynie program FIEFdom osiągnął zbliżoną skuteczność, ale ciągle DomainSVM był lepszy o blisko 3%. Ponadto oba wspomniane programy używają informacji o lokalizacji domen w homologicznych białkach. Jak wspomniano wyżej, przy jej braku skuteczność DomainSVM spada; jednak wersja pozbawiona tej cechy jest lepsza niż najlepszy klasyfikator o podobnej charakterystyce (DomPred (DGS)) – tabela 11.

W przypadku w którym program zakwalifikuje daną sekwencję jako należącą do białka wielodomenowego, następuje przejście do drugiej warstwy, której zadanie polega na właściwym przewidzeniu regionów odpowiadających granicom domen. Klasyfikator tej warstwy oparty jest na cechach podobnych jak w przypadku pierwszej warstwy, z tym, że wektor reprezentuje pojedynczą resztę aminokwasową zamiast całej sekwencji. Ponadto dodano tu dodatkowo cechy takie jak:

- relatywna pozycja reszty aminokwasowej w sekwencji (pozycja reszty podzielona przez długość sekwencji),
- odległość od końca N i odległość od końca C,
- przewidywana liczba kontaktów dla pięcioaminokwasowego okna wyliczona według macierzy zdefiniowanej w (Miyazawa i Jernigan, 1999),
- macierz PSSM wyliczona za pomocą programu PSI-BLAST.

Podobnie jak poprzednio program porównano z innymi, jednak tym razem różnica między DomainSVM a FIEFdom nie jest tak duża jak w przypadku testów pierwszej warstwy (tabela 12).

Podsumowując, należy stwierdzić, że zastosowanie rozwiązania polegającego na podzieleniu problemu na dwa mniejsze znacząco podnosi skuteczność programu, tj. zmniejsza ono liczbę wyników fałszywie pozytywnych, czyli granic domen przewidzianych w obrębie obszarów pozagranicznych.

7.3. Białka odpowiedzialne za modyfikację końca 3′ mRNA – biologiczny przykład zastosowania programów do przewidywania domen i regionów wewnętrznie nieuporządkowanych

Wszystkie przedstawione w niniejszej rozprawie programy mają za zadanie pomóc w weryfikowaniu hipotez badawczych postawionych przez biologów. W ramach przedstawionych badań dokonano analizy ludzkich białek odpowiedzialnych za modyfikację końca 3′ mRNA. Jako pierwszy cel badawczy autor postawił sobie za zadanie zebranie aktualnie dostępnej informacji na temat tej grupy białek. Następnie za pomocą programów zaprezentowanych w niniejszej rozprawie dokonał analizy ze szczególnym uwzględnieniem występowania wewnętrznego nieuporządkowania i domen. Wyniki tej części badań udostępnione zostały w formie internetowej bazy danych (http://www.genesilico.pl/mrna3db/). Zawiera ona zarówno dane zaczerpnięte z literatury jak i wyniki opisane poniżej (przykładowa strona pochodząca z bazy dotycząca białka PABP-1 została zaprezentowana na ryc. 20).

7.3.1. Regiony wewnętrznie nieuporządkowane i ich znaczenie

Jednym z najważniejszych wniosków wynikających z badania białek odpowiedzialnych za modyfikację końca 3′ mRNA jest wysoka częstotliwość występowania regionów klasyfikowanych jako wewnętrznie nieuporządkowane przez program MetaDisorderMD2; średnio około 50% reszt klasyfikowanych jest jako wewnętrznie nieuporządkowane (tabela 14), analogicznie wartość ta dla całego proteomu wynosi niewiele ponad 20%. Ponadto około 80% białek posiada przynajmniej jeden region IUR dłuższy niż 30 reszt aminokwasowych. Ogólnie można powiedzieć, że białka rozpatrywanego kompleksu są duże (średnia długość powyżej 600 reszt) i wraz ze wzrostem długości wzrasta liczba reszt klasyfikowanych jako wewnętrznie nieuporządkowane. Wskazuje to jednoznacznie, że regiony wewnętrznie nieuporządkowane są niezwykle istotne dla funkcjonowania kompleksu. Potwierdza to każe fakt, że białka kompleksu

konserwowane są na przestrzeni ewolucji tj. większość ludzkich białek posiada swoje homologi w białkach drożdży). Otrzymany wynik zdaje się potwierdzać ostatnie badania, według których białka oddziałujące z mRNA, na wszystkich etapach jego trwania, cechują się wysokim stopniem nieuporządkowania (Castello et al., 2012). Ponadto przewidywana liczba regionów IUR jest bardzo podobna do opisanej dla białek ludzkiego spliceosomu (Korneta i Bujnicki, 2012; Korneta et al., 2012), innego wielobiałkowego, dynamicznego kompleksu oddziałującego z mRNA. Dodatkowo, w czasie badań dokonano analizy poszczególnych regionów IUR pod kątem takich cech jak złożoność sekwencji lub/i potencjalna struktura drugorzędowa, (interaktywna wersja tabeli 21 pozwalająca na sortowanie po kolumnach dostępna jest na stronie internetowej bazy danych na podstronie „Disorder"). Dane pozwalają na łatwe znajdywanie potencjalne ciekawych regiony IUR, np. posiadających wysoki odsetek struktury drugorzędowej (region 602-645 w białku PABP-4) co sugeruje, że mogą one prawdopodobnie ulegać ustrukturalizowaniu (także ryc. 24).

7.3.2. Domeny strukturalne w białkach kompleksu odpowiedzialnego za modyfikację końca 3′ mRNA

Kolejnym ważnym aspektem analizy białek odpowiedzialnych za modyfikację końca 3′ mRNA poruszonym w niniejszej pracy jest badanie składu domen oraz przewidywanie ich granic. Wśród znanych domen oznaczonych za pomocą bazy PFAM trzy najczęściej występujące grupy domen to domeny RRM, domeny WD40 oraz rożne typy motywu palca cynowego. Wynik nie dziwi, ponieważ wspomniane domeny odpowiadają za wiązanie RNA i tworzenie oddziaływań białek z innymi białkami. Najciekawszym wynikiem tej części analizy jest stwierdzenie, że częstotliwość przewidywanych elementów struktury drugorzędowej jest niemal identyczna dla obszarów domenowych i obszarów pozadomenowych (tabela 16). Dodatkowo, biorąc pod uwagę procentową zawartość regionów IUR dane wskazują, że regiony wewnętrznie nieuporządkowane posiadają nadzwyczajnie dużo potencjalnych regionów, które wykazują tendencje do tworzenia helis α i wstęg β. Wynik ten jest nieoczekiwany i zdaje się potwierdzać hipotezę, według której regiony IUR nie dość, że są istotne, to prawdopodobnie jeszcze spora część z nich ulega ustrukturalizowaniu pod wpływem innych białek i mRNA, przyjmując ścisłe określoną strukturę zależną od partnera.

7.3.3. Model białka CPSF3 przykładem możliwości zastosowania modelowania homologicznego.

Jednym z ważniejszych wyników analizy białek omawianego kompleksu było zbudowanie modeli homologicznych w oparciu o szablony wykryte przez metaserwer GeneSilico. Dla każdego z białek w automatyczny sposób zbudowano około 100 modeli następnie na podstawie jakości modelu (ocenionej za pomocą programu MetaMQAPII), wybrano najlepsze z nich. Jako ciekawy przykład możliwości modelowania homologicznego zaprezentowano model białka CPSF3. Białko posiada funkcję endonukleolityczną i odpowiada za cięcie mRNA w specyficznym miejscu. O ile sama domena endonukleazy (region 1-459) znajdująca się po stronie N-końcowej została rozwiązana, o tyle struktura odpowiadająca około 200 resztom od strony końca N nie jest znana. Wnioskując po wynikach przewidywania obszarów wewnętrznie nieuporządkowanych, przynajmniej połowa tego regionu jest uporządkowana. Należy jednak zwrócić uwagę, że również przewidywanie struktury drugorzędowej wskazuje na lokalizację w tym regionie helis α i wstęg β. Model, który został wygenerowany na podstawie szablonu 3ZQ4 dobudował w tym regionie domenę składającą się z trzech wstęg β i dwóch helis α (ryc. 24), która kształtem przypomina domenę RRM (ryc. 23a). Funkcja modelowanej domeny zgadza się potencjalną funkcją białka (wiązanie RNA), przykład ten może być przykładem regionu wewnętrznie nieuporządkowanego, który ustrukturalizowuje się przy interakcji z mRNA.

7.4. Metaserver GeneSilico jako przykład dużego projektu bioinformatycznego

W chwili obecnej autor rozprawy jest osobą odpowiedzialną za utrzymanie i rozwój metaserwera GeneSilico. Jest on udostępniony w formie bezpłatnego serwisu internetowego, który w łatwy sposób pozwala uruchomić ponad 120 programów bioinformatycznych (tabela 2, bardziej szczegółowy opis metod wchodzących w skład metaserwera znajduje się w rozdziale 5.3.4.). Choć praca opisująca ten serwis została opublikowana w 2003 roku (Kurowski i Bujnicki, 2003), obecny jego stan znacząco odbiega od pierwowzoru. Po pierwsze, serwis został znacząco rozbudowany poprzez dodanie nowych programów (w 2003 roku było to zaledwie 17 programów bioinformatycznych, z których większość służyła do rozpoznawania zwoju białka).

Aktualnie metaserwer pozwala na przewidywanie takich cech jak struktura drugorzędowa, obecność helis transbłonowych, sekwencji sygnalnych, mostków dwusiarczkowych i struktur splecionych helis, dostępność reszt aminokwasowych dla rozpuszczalnika, domeny oraz wewnętrzne nieuporządkowanie. Ponadto metaserwer został zintegrowany z kilkoma innymi programami, umożliwiając przeprowadzanie dalszej analizy pod kątem specjalistycznych cech (np. weryfikacji użyteczności modeli homologicznych dla techniki podstawienia cząsteczkowego za pomocą programu AmIgoMR (Pawlowski i Bujnicki, 2012)).

Metaserwer GeneSilico jest unikatowym serwisem bioinformatycznym na skalę światową. Można go jedynie porównać do dwóch tego typu serwisów, a mianowicie do serwisu PredictProtein (Rost i Liu, 2003) oraz do niedawna do metaserwera BioInfoBanku (Bujnicki et al., 2001), który został wyłączony w sierpniu 2012 z powodu braku finansowania ze źródeł publicznych. Nawet jeśli pominąć liczbę programów, które dany serwis może uruchomić (metaserwer GeneSilico pod tym względem jest bezkonkurencyjny), zdecydowanie przewyższa on funkcjonalnością oba wymienione. Podstawowa różnica między metaserwerem a portalem PredictProtein, polega na tym, że ten ostatni pozwala jedynie na uruchomienie określonej liczby programów, a wyniki nie są przetwarzane (użytkownik otrzymuje listę wyników w niejednorodnym formacie, który znacząco utrudnia interpretacje wyników). Na tym tle metaserwer wypada znacznie lepiej, wyniki programów poszczególnych kategorii przetwarzane są do intuicyjnego formatu, a następnie przyrównane zostają do sekwencji-celu (ryc. 14).

Ogólnie architekturę metaserwera można podzielić na trzy warstwy (ryc. 10). Pierwszą z nich jest warstwa zewnętrzna widoczna dla użytkownika (została ona napisana w serwerze aplikacji Zope) w formie interfejsu internetowego. Pozwala ona na wygodny dostęp do indywidualnych wyników. Następnie można wyróżnić warstwę zbudowaną z baz danych, w której przechowywane są wszelkie informacje (przewidywania programów w oryginalnym formacie, przyrównania do szablonów oraz modele homologiczne). Warstwę niejako leżącą na samym dnie stanowią skrypty napisane w języku Python, które odpowiadają za uruchamianie i przetwarzanie wyników programów wchodzących w skład metaserwera. Ponadto do warstwy tej należy zaliczyć inne skrypty odpowiedzialne za funkcje pomocnicze (kolejkowanie zapytań, aktualizacja biologicznych baz danych itd.). Wraz z rozwojem serwisu niezbędne było prawie całkowite przebudowanie architektury w stosunku do pierwotnej wersji serwisu. Przykładowo skrypty zarządzające działaniem serwisu były napisane w języku Perl, który okazał się na tyle

niewygodny w użyciu, że niezbędne było ich przepisanie na język Python. Wiele innych rozwiązań było zmieniane w momencie, gdy przestawały one być użyteczne. Tu jako przykład można podać system kolejkowania, który początkowo opierał się na systemie plików, które zawierały status kolejki. Niestety rozwiązanie to okazało się mało wydajne w momencie dużego obciążenia komputera, na którym zainstalowany był metaserwer. W takich przypadkach skrypt kolejkujący, czytając dane z dysku komputera, musiał „walczyć" o zasoby z programami uruchamianymi w tle przez metaserwer. Powodowało to komplikacje takie jak niemożność zapisania lub odczytania aktualnego stanu kolejki lub w skrajnym przypadku wyczyszczenie plików kolejki. Problem ten rozwiązano, przenosząc kolejkowanie z plików na dysku do bazy danych.

Ponadto, wraz z rozwojem metaserwera okazało się, że nie ma wydajnej metody na przewidzenie liczby zasobów, która zostanie wykorzystana w danym momencie przez kilkadziesiąt programów. Sytuacje, w których dochodzi do przeciążenia systemu są nieuniknione. Wynika to z faktu, że zużycie mocy obliczeniowej i pamięci operacyjnej RAM są różne dla każdego zainstalowanego programu i zmieniają się, czasem w sposób niemożliwy do przewidzenia, wraz z długością sekwencji. Wpływ może mieć tu nawet skład aminokwasowy sekwencji, która szybciej lub wolniej zostanie odnaleziona w bazie sekwencji homologicznych. Mając to na uwadze, postanowiono rozgraniczyć dwie pierwsze warstwy od skryptów uruchamiających programy składowe. W ten sposób warstwę interfejsu użytkownika i bazę danych umieszczono na oddzielnej, względnie słabej maszynie, natomiast warstwę skryptów i wszystkie programy, które przez nią są uruchamiane ulokowano na innym wydajniejszym serwerze (48 CPU, 96 GB RAM). Rozgraniczając warstwę odpowiadającą za prezentację danych oraz same dane od warstwy obliczeniowej, serwis zyskał większą stabilność (wgląd do wyników na stronie internetowej jest zawsze możliwy bez względu na to jak bardzo obciążony obliczeniowo zostanie metaserwer). Już sama liczba programów, które uruchamia metaserwer powoduje, że niezbędna jest duża moc obliczeniowa, dodatkowo duża część programów działa w oparciu o różne biologiczne bazy danych (rozdział 5.2), które co jakiś czas muszą być aktualizowane. Proces ten może być wymagający zarówno pod względem miejsca na dysku (np. lokalna wersja bazy danych PDB zawiera ponad 80 GB danych) jak i mocy obliczeniowej (np. baza nr90 tworzona jest co tydzień na podstawie bazy nr poprzez uruchomienie programu cd-hit, całość trwa ponad 16 godzin, nawet jeśli wykorzystane zostanie wszystkie 48 procesorów).

Najważniejszym usprawnieniem, jakiego dokonał autor niniejszej rozprawy, która znacząco wpłynęła na stabilność metaserwera i umożliwiła jego dalszy rozwój było zaprojektowanie wydajnej struktury zarządzającej uruchomieniem programów wchodzących w skład metaserwera. W chwili obecnej pojedynczy program zarządzany jest przez oddzielny, w pełni niezależny moduł (jego szczegółowy opis został podany w rozdziale 6.1.4). W ten sposób, problemy z jednym programem nie blokują działania innych. W przypadku programów lokalnie zainstalowanych, pomimo opisanych wcześniej problemów, sytuację i tak należy uznać za komfortową. Zwykle raz zainstalowany program nie przysparza problemów w przyszłości. W skład metaserwera wchodzą także zewnętrzne serwisy internetowe. Tu lista możliwych komplikacji jest bardzo długa. Właściwie wszystko może się wydarzyć. Poczynając od wyłączenia serwisu, brak odesłania wiadomości na wskazany email, a kończąc na zmianie formatu formularza lub strony stanowiącej odpowiedz serwisu. Aby uniknąć przynajmniej części z tych problemów, w metaserwer wbudowany został system pozwalający na ponowienie wysłania sekwencji jeśli metoda nie zwróci wyniku po ustalonym czasie. Niestety trzeba tu wspomnieć, że średni czas „życia" serwisów internetowych jest stosunkowo krótki i zwykle po 2-3 latach, gdy autor programu opuści laboratorium przestają one działać. W tym momencie cały trud związany z integracja takiej metody zostaje zaprzepaszczony. Ogólnie można powiedzieć, że na każdy program, którego aktualnie stanowi składnik metaserwera przypada drugi, który przestał działać i musiał być usunięty.

7.5. Celowość wykorzystania meta-metod

Wyniki prezentowane w niniejszej rozprawie pokazują, że stosując meta-metody, można łatwo otrzymać istotne wyniki badawcze tak jak to było w przypadku analizy ludzkich białek odpowiedzialnych za modyfikację końca 3′ mRNA. Z jednej strony użytkownicy mogą łatwo weryfikować nawet złożone hipotezy badawcze, zaś z drugiej konstrukcja meta-metody pozwala na wykorzystanie aktualnego stanu wiedzy reprezentowanej przez programy wchodzące w skład metametody. Poszczególne programy można traktować jako niezależne bloki, które reprezentują wiedzę autora, wykorzystany zbiór danych treningowych oraz technikę uczenia maszynowego. Oczywiście można podejść do sprawy w tradycyjny sposób i po sprawdzeniu jaki komponent decyduje o skuteczności danego programu (np. zapoznając się z publikacją opisującą dany

program) próbować dodać taką cechę jako składnik naszego programu. Z drugiej strony, chyba nie ma takiej potrzeby. Znacznie wydajniejszym podejściem wydaje się rozbudowanie już istniejącego programu o nową funkcjonalność (tak postąpiono w przypadku dodania do programu FloatCons informacji o występowaniu przerw w dopasowaniach co przyczyniło się do powstania lepszej metody MetaDisorderMD2).

Choć przeciwnicy meta-metod starają się przedstawić je jako programy „pasożytnicze", które nie stymulują do dalszego rozwoju programów danego typu przykład programu MetaDisorder jednoznacznie temu przeczy. Na bazie istniejących programów powstał skuteczniejszy program do przewidywania regionów wewnętrznie nieuporządkowanych, a zaraz po jego sukcesie w eksperymencie CASP8 powstała cała grupa programów naśladujących wykorzystany przez autora model (m.in. programy MetaPrdos (Ji et al., 2013), PONDR-FIT (Xue et al., 2010)).

7.6. Statystyki użytkowania serwisów

W rozprawie doktorskiej wspomniano kilkakrotnie, że metaserwer GeneSilico jest serwisem internetowym, z którego korzysta wielu badaczy z całego świata. Na ryc. 25 pokazano statystyki odwiedzin użytkowników z ostatnich pięciu lat. W styczniu 2013 roku liczba zarejestrowanych użytkowników metaserwera GeneSilico przekroczyła 2300 (w 2008 roku w momencie w którym autor rozprawy przejął obowiązki związane z rozwojem serwisu użytkowników było zaledwie 500). Tylko niewielka liczba użytkowników pochodzi z Polski lub/i macierzystego laboratorium (odpowiednio 408 i 57 użytkowników). Pozostali użytkownicy najczęściej pochodzą z Stanów Zjednoczonych, Indii, Hiszpanii, Wielkiej Brytanii i Niemiec (odpowiednio 366, 360, 181, 126, 105 użytkowników). Także liczba zapytań, analizowanych sekwencji stabilnie wzrasta (o ile do roku 2008 zapytań tych było niecałe 12 tysięcy, to w lutym 2013 roku liczba ta przekroczyła 33 tysiące). Należy podkreślić, że pojedyncze zapytanie (wysłanie jednej tylko sekwencji) może pozwolić na weryfikację nawet dość skomplikowanej hipotezy badawczej (metaserwer na podstawie wysłanej sekwencji generuje olbrzymią ilość informacji, np. przewidywania wielu cech białka, dziesiątki przyrównań i modeli homologicznych).

Mimo że od publikacji oryginalnej pracy opisującej metaserwer (Kurowski i Bujnicki, 2003) upłynęło ponad 10 lat, dzięki ustawicznemu ulepszaniu serwisu (zarówno pod kątem liczby metod składowych jak i poprawy funkcjonalności interfejsu) praca ta jest najczęściej cytowaną pracą pochodzącą z macierzystego laboratorium i aktualnie ma ponad 200 cytowań (ponad 30 cytowań w samym tylko roku 2012). W najbliższym czasie planowana jest aktualizacja publikacji mająca na celu opisanie obecnej wersji metaserwera GeneSilico (manuskrypt w przygotowaniu).

Na tym tle MetaDisorder, drugi z serwisów prezentowanych w niniejszej rozprawie, jest serwisem względnie młodym (został uruchomiony w 2012 roku), lecz mimo to skorzystano z niego ponad 2200 razy.

Ryc. 25. Statystyki użytkowania metaserwera GeneSilico. Na górze pokazano mapę obrazującą źródła odwiedzin w zależności od pochodzenia użytkowników. Na dole pokazano analogiczne dane w formie wykresu uwzględniającego czas wizyty. Za zbieranie statystyk odpowiedzialny jest serwis Google Analitics.

7.7. Implementacja metod, upublicznienie wyników badań

W ramach niniejszego projektu powstały metody do przewidywania wewnętrznego nieuporządkowania (MetaDisorder) oraz do przewidywania domen w białkach (DomainSVM). Dodatkowo znacząco rozbudowany został metaserwer GeneSilico (https://genesilico.pl/meta2/). Badania dotyczące wewnętrznego nieuporządkowania zostały opublikowane w czasopiśmie BMC Bioinformatics (Kozlowski i Bujnicki, 2012), metoda MetaDisorder dostępna jest również jako bezpłatny serwis internetowy (http://genesilico.pl/metadisorder/). Dodatkowo została ona zintegrowana z metaserwerem GeneSilico. Program DomainSVM także dostępny jest z poziomu metaserwera GeneSilico (obecnie trwają prace nad manuskryptem oraz implementacją tej metody jako niezależnego serwisu internetowego).

Dodatkowym wynikiem prezentowanych badań jest internetowa baza danych dotycząca ludzkich białek odpowiedzialnych za modyfikację końca 3′ mRNA (http://www.genesilico.pl/mrna3db/). Zawiera ona informacje odzwierciedlające aktualny stan wiedzy dotyczący rozpatrywanej grupy białek oraz wyniki analizy zaprezentowanej w rozprawie.

7.8. Przyszłe kierunki badań

W przypadku metaserwera GeneSilico oczywistym wydaje się, że rozwój serwisu może przebiegać na dwa sposoby. Po pierwsze, poprzez dodanie nowych kategorii programów np. przewidujących mapy kontaktów (w chwili obecnej na etapie testów) bądź miejsc posttranslacyjnej modyfikacji. Po drugie, poprzez udoskonalenie obecnego interfejsu użytkownika (np. możliwość wyboru określonej kategorii programów, w chwili obecnej domyślnie uruchamiane są programy wszystkich kategorii). Wszystkie te zmiany są niezwykle pracochłonne i ich przeprowadzanie wymaga dużej ilości czasu.

W niniejszej rozprawie zaprezentowano także programy MetaDisorder i DomainSVM i, choć jak jednoznacznie wykazano, należą one do grona najlepszych programów w swojej kategorii, ciągle istnieje możliwość ich ulepszenia.

Przewidywanie wewnętrznego nieuporządkowania w białkach można prawdopodobnie ulepszyć na kilka sposobów.

1. W chwili obecnej programy przewidujące regiony IUR w pierwszym etapie przeszukują duże biologiczne bazy danych (np. nr) w celu zbudowania odpowiadającego im profilu. Podejście to jest ogólne słuszne, ale uwzględniając specyfikę problemu bardziej celowe wydaje się wykorzystanie bazy ukierunkowanej na białka wewnętrznie nieuporządkowane. Choć baza DisProt posiada funkcjonalność polegającą na jej przeszukaniu za pomocą programu RPS-BLAST, to według wiedzy autora nikt jeszcze nie próbował zintegrować przeszukiwania bazy DisProt z klasycznym klasyfikatorem opartym na metodach uczenia maszynowego. Ponadto, innym możliwym usprawnieniem mogłoby być zastosowanie modeli HMM zamiast profili PSSM, ponieważ zwykle dają one lepszy wynik.

2. Jednym z podstawowych ograniczeń obecnej wersji programu MetaDisorder jest nierozróżnianie poszczególnych klas wewnętrznego nieuporządkowania. Omawiając wyniki analizy białek odpowiedzialnych za modyfikację końca 3′ mRNA, pokazano jak za pomocą innych cech białka, takich jak przewidywana struktura drugorzędowa czy poziom złożoności sekwencji można podzielić regiony wewnętrznie nieuporządkowane na oddzielne klasy. Wyniki tej analizy okazały się na tyle obiecujące, że następnym krokiem mogłoby być zautomatyzowanie zastosowanej metodologii. Przykładowo, w chwili obecnej MetaDisorder pozwala na analizę tylko jednej sekwencji. Dodanie prostej funkcjonalności polegającej na umożliwieniu jednoczesnej analizy większej liczby białek oraz porównaniu otrzymanych wyników do przewidywania struktury drugorzędowej, poziomu złożoności sekwencji i innych cech białka byłoby korzystne.

3. Innym aspektem, który nie jest w tej chwili brany pod uwagę, jest otoczenie w jakim znajduje się białko. Jak wielokrotnie podkreślano w niniejszej rozprawie, przynajmniej część regionów wewnętrznie nieuporządkowane zmienia swoją konformację pod wpływem innych białek. Niestety o procesie tym ciągle prawie nic nie wiadomo. Ciekawym pytaniem jest, więc czy istnieje możliwość skonstruowania programu, który dla dwóch lub większej liczby białek byłby w stanie symulować proces interakcji regionów wewnętrznie nieuporządkowanych z konkretnym regionem innej makrocząsteczki. Wydaje się, że dane, które można by wykorzystać do rozwiązania tego

problemu już są dostępne (np. baza IDEAL, w której gromadzone są dane literaturowe opisujące znane przypadki regionów IUR, które cechują się przejściem ze stanu nieuporządkowania w stan zdefiniowanej struktury trzeciorzędowej, indukowanym przez oddziaływanie z białkiem).

4. Kolejnym, czysto technicznym usprawnieniem mogłoby być wykorzystanie innych, lepszych technik uczenia maszynowego bądź po prostu ponowne przetrenowanie metody na nowych zbiorach sekwencji (przykładowo program MetaDisorder trenowany był na wersji bazy DisProt z 2008 roku, od tego czasu liczba sekwencji w tej bazie wzrosła o połowę).

Także w przypadku programu DomainSVM istnieje kilka możliwości na poprawienie działania. Przykładowo w chwili obecnej działanie programu ogranicza się do przewidywania granic domen, jednak pod kątem analizy otrzymanych wyników istotna jest także identyfikacja rodzaju domeny. W najprostszym wariancie za etap ten mógłby odpowiadać program HMMER, który na podstawie fragmentów powstałych w oparciu o granice przewidziane przez program DomainSVM sprawdzałby podobieństwo do znanych domen. Z jednej strony użytkownik dostawałby informację o występowaniu konkretnej domeny, zaś z drugiej pozwoliłoby to wskazać domeny na temat, których nie ma żadnych informacji, wskazując je jako interesujący obiekt badawczy. Innym problemem w przypadku programu DomainSVM jest użycie tzw. „płaskiej" definicji domeny, która ogranicza użyteczność programu do domen ciągłych zlokalizowanych w jednym białku. Zastosowanie definicji która obejmowałaby domeny nieciągłe i domeny złożone z innych łańcuchów (ang. *domain swapping*) wymaga skonstruowania innego, bardziej skomplikowanego zbioru uczącego co wcale nie oznacza, ze programów o takich możliwościach nie ma (np. program 3dswap-pred (Shameer et al., 2011a)).

Część proponowanych pomysłów powstała w trakcie analizy białek odpowiedzialnych za modyfikację końca 3′ mRNA. Analiza ta dotyczyła białek ludzkich. Oczywistym kierunkiem tej część badań mogłaby być podobna analiza, ale dotycząca białek drożdży (także w ich przypadku znane są poszczególne składniki kompleksu odpowiedzialnego za ten proces) lub też analiza kompleksu o podobnej funkcji, który odpowiada za modyfikację końca 3′ mRNA transkryptów

białek histonowych (kompleksy te różnią się znacząco pod względem składu poszczególnych białek).

8. Podsumowanie

Mimo, iż w chwili obecnej znamy blisko 90 tysięcy struktur białkowych, liczba ta jest stosunkowo niewielka, jeśli weźmie się pod uwagę liczbę znanych sekwencji białkowych, która przekroczyła 20 milionów. Ponadto rozwiązanie struktury białek metodami doświadczalnymi jest procesem żmudnym i kosztownym, a na dodatek jest obarczone wysokim ryzykiem niepowodzenia. Jedynie wspomagając się modelowaniem komputerowym możemy poprawić istniejącą sytuację.

W ramach niniejszej rozprawy zaprezentowano zintegrowany serwis bioinformatyczny do analizy białek. Pozwala on w łatwy sposób uruchomić ponad 100 programów przewidujących cechy takie jak ogólny zwój białka, jego struktura drugorzędowa, obecność helis transbłonowych, sekwencji sygnalnych, mostków dwusiarczkowych i struktur splecionych helis, dostępność reszt aminokwasowych dla rozpuszczalnika, domeny oraz wewnętrzne nieuporządkowanie. Wyniki poszczególnych klas programów przedstawione są w intuicyjnym formacie, który pozwala łatwo porównać ich wyniki.

Dodatkowo na bazie wspomnianego serwisu stworzony został nowy program do przewidywania regionów wewnętrzne nieuporządkowanych. Program ten jest meta-metodą, czyli agreguje wyniki innych programów i po ich przetworzeniu za pomocą metod uczenia maszynowego zwraca potencjalnie ulepszony wynik. Zarówno testy przeprowadzone przez autora niniejszej pracy jak i niezależne testy w ramach międzynarodowego eksperymentu CASP potwierdziły jego użyteczność (program MetaDisorder był najlepszym programem w swojej kategorii w przeciągu dwóch kolejnych edycji konkursu tj. w czasie CASP8 i CASP9). Wynik ten wykazuje, że użycie meta-metodologii do konstruowania nowych programów do przewidywania wewnętrznego nieuporządkowania w białkach jest uzasadnione i skuteczne.

Kolejnym problemem poruszonym w niniejszej rozprawie jest przewidywanie domen. Poprawna identyfikacja domen w obrębie sekwencji analizowanego białka jest bardzo ważna, ponieważ wpływa ona na jakość przewidywania innych programów (zwykle działają one dobrze jedynie dla białek jednodomenowych) oraz na skuteczność technik doświadczalnych (bardzo często nie udaje się rozwiązać struktury całego białka, jednak jest to możliwe dla poszczególnych jego domen). W tym celu stworzono program DomainSVM. Zastosowano maszynę wektorów nośnych (SVM), której działanie uwzględniało cechy takie jak: entropia, hydrofobowość, przewidywana struktura drugorzędowa, dostępność reszt aminokwasowych dla

rozpuszczalnika, występowanie regionów wewnętrznie nieuporządkowanych oraz obecność bliskich homologów w bazie domen białkowych CATH. W efekcie powstał program, który według przeprowadzonych testów daje lepsze wyniki niż inne tego typu programy. Należy podkreślić, że główną cechą odpowiedzialną za skuteczność programu jest informacja dotycząca domen pochodząca z białek homologicznych. Jednak nawet w przypadku braku tej informacji program przewiduje granice domen z akceptowalną skutecznością.

Oprócz testów statystycznych prezentowane narzędzia bioinformatyczne przetestowano w sposób praktyczny. Za ich pomocą dokonano analizy białek odpowiedzialnych za modyfikację końca 3′ mRNA. Korzystając z metaserwera GeneSilico zidentyfikowano najlepsze szablony i na ich podstawie zbudowano modele homologiczne. Za pomocą programów MetaDisorder i DomainSVM przewidziano występowanie regionów wewnętrznie nieuporządkowanych oraz granic domen. Najważniejszym wnioskiem z tej części badań było stwierdzenie faktu, że regiony wewnętrznie nieuporządkowane stanowią około połowy długości białek odpowiedzialnych za modyfikację końca 3′ mRNA. Zauważono ogólną tendencję, według której dłuższe białka posiadają więcej regionów nieuporządkowanych. Ponadto, wykazano, że regiony wewnętrznie nieuporządkowane kompleksu posiadają nadzwyczajnie dużo regionów, które wykazują tendencje do tworzenia helis α i wstęg β. Wynik ten jest dość nieoczekiwany i zdaje się potwierdzać hipotezę, według której regiony IUR ulegają ustrukturalizowaniu pod wpływem innych białek i mRNA, przyjmując ścisłe określoną strukturę zależną od partnera.

9. Bibliografia

Adamczak, R., Porollo, A., i Meller, J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. Proteins *59*, 467-475.

Alexandrov, N., i Shindyalov, I. (2003). PDP: protein domain parser. Bioinformatics *19*, 429-430.

Almen, M.S., Nordstrom, K.J., Fredriksson, R., i Schioth, H.B. (2009). Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biol *7*, 50.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., i Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science *181*, 223-230.

Anfinsen, C.B., Haber, E., Sela, M., i White, F.H., Jr. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci U S A *47*, 1309-1314.

Bennett-Lovsey, R.M., Herbert, A.D., Sternberg, M.J., i Kelley, L.A. (2007). Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins.

Berg, J.M. (2007). Biochemistry, 6th ed. edn (New York :, W. H. Freeman).

Berger, B., Wilson, D.B., Wolf, E., Tonchev, T., Milla, M., i Kim, P.S. (1995). Predicting coiled coils by use of pairwise residue correlations. Proc Natl Acad Sci U S A *92*, 8259-8263.

Bernsel, A., Viklund, H., Falk, J., Lindahl, E., von Heijne, G., i Elofsson, A. (2008). Prediction of membrane-protein topology from first principles. Proc Natl Acad Sci U S A *105*, 7177-7181.

Bernsel, A., Viklund, H., Hennerdal, A., i Elofsson, A. (2009). TOPCONS: consensus prediction of membrane protein topology. Nucleic Acids Res *37*, W465-468.

Bettella, F., Rasinski, D., i Knapp, E.W. (2012). Protein secondary structure prediction with SPARROW. J Chem Inf Model *52*, 545-556.

Biegert, A., i Soding, J. (2009). Sequence context-specific profiles for homology searching. Proc Natl Acad Sci U S A *106*, 3770-3775.

Bondugula, R., Lee, M.S., i Wallqvist, A. (2009). FIEFDom: a transparent domain boundary recognition system using a fuzzy mean operator. Nucleic Acids Res *37*, 452-462.

Bujnicki, J.M., Elofsson, A., Fischer, D., i Rychlewski, L. (2001). Structure prediction meta

server. Bioinformatics *17*, 750-751.

Butterfield, A., Vedagiri, V., Lang, E., Lawrence, C., Wakefield, M.J., Isaev, A., i Huttley, G.A. (2004). PyEvolve: a toolkit for statistical modelling of molecular evolution. BMC Bioinformatics *5*, 1.

Bystroff, C., Thorsson, V., i Baker, D. (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol *301*, 173-190.

Cai, Y.D., i Lu, L. (2008). Predicting N-terminal acetylation based on feature selection method. Biochem Biophys Res Commun *372*, 862-865.

Cao, B., Porollo, A., Adamczak, R., Jarrell, M., i Meller, J. (2006). Enhanced recognition of protein transmembrane domains with prediction-based structural profiles. Bioinformatics *22*, 303-309.

Carpenter, J., i Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in medicine *19*, 1141-1164.

Carson, M.B., Langlois, R., i Lu, H. (2010). NAPS: a residue-level nucleic acid-binding prediction server. Nucleic Acids Res *38*, W431-435.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M.*, et al.* (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell *149*, 1393-1406.

Ceroni, A., Passerini, A., Vullo, A., i Frasconi, P. (2006). DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. Nucleic Acids Res *34*, W177-181.

Chang, C.-C., i Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology *2*, 27:21-27:27.

Cheng, H., Sen, T.Z., Jernigan, R.L., i Kloczkowski, A. (2007). Consensus Data Mining (CDM) Protein Secondary Structure Prediction Server: combining GOR V and Fragment Database Mining (FDM). Bioinformatics *23*, 2628-2630.

Cheng, H., Sen, T.Z., Kloczkowski, A., Margaritis, D., i Jernigan, R.L. (2005a). Prediction of protein secondary structure by mining structural fragment database. Polymer (Guildf) *46*, 4314-4321.

Cheng, J. (2007). DOMAC: an accurate, hybrid protein domain prediction server. Nucleic Acids Res *35*, W354-356.

Cheng, J., Saigo, H., i Baldi, P. (2006a). Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. Proteins *62*, 617-629.

Cheng, J., Sweredoski, M., i Baldi, P. (2005b). Accurate prediction of protein disordered regions

by mining protein structure data. Data Mining and Knowledge Discovery *11*, 213-222.

Cheng, J., Sweredoski, M.J., i Baldi, P. (2006b). DOMpro: Protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. Data Mining and Knowledge Discovery *13*, 1-10.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. J Mol Biol *105*, 1-12.

Cole, C., Barber, J.D., i Barton, G.J. (2008). The Jpred 3 secondary structure prediction server. Nucleic Acids Res *36*, W197-201.

Cserzo, M., Eisenhaber, F., Eisenhaber, B., i Simon, I. (2004). TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics *20*, 136-137.

Danckwardt, S., Hentze, M.W., i Kulozik, A.E. (2008). 3' end mRNA processing: molecular mechanisms and implications for health and disease. Embo J *27*, 482-498.

Darmon, S.K., i Lutz, C.S. (2012). mRNA 3' end processing factors: a phylogenetic comparison. Comp Funct Genomics *2012*, 876893.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System (San Carlos, CA, USA, DeLano Scientific).

Denning, D.P., Patel, S.S., Uversky, V., Fink, A.L., i Rexach, M. (2003). Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. Proc Natl Acad Sci U S A *100*, 2450-2455.

Di Domenico, T., Walsh, I., Martin, A.J., i Tosatto, S.C. (2012). MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics *28*, 2080-2081.

Dor, O., i Zhou, Y. (2007). Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins *66*, 838-845.

Dosztanyi, Z., Csizmok, V., Tompa, P., i Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433-3434.

Dosztanyi, Z., Sandor, M., Tompa, P., i Simon, I. (2007). Prediction of protein disorder at the domain level. Curr Protein Pept Sci *8*, 161-171.

Dunbrack, R.L., Jr. (1999). Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. Proteins *Suppl 3*, 81-87.

Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., i Uversky, V.N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. Febs J *272*, 5129-5148.

Egloff, S., i Murphy, S. (2008). Cracking the RNA polymerase II CTD code. Trends in genetics :

TIG *24*, 280-288.

Eickholt, J., Deng, X., i Cheng, J. (2011). DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning. BMC Bioinformatics *12*, 43.

Ellis, R.J. (2006). Molecular chaperones: assisting assembly in addition to folding. Trends Biochem Sci *31*, 395-401.

Ezkurdia, I., i Tress, M.L. (2011). Protein structural domains: definition and prediction. Current protocols in protein science / editorial board, John E Coligan [et al] *Chapter 2*, Unit2 14.

Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., i Zhou, Y. (2012). SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J Comput Chem *33*, 259-267.

Ferre, F., i Clote, P. (2006). DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine classification. Nucleic Acids Res *34*, W182-185.

Finn, R.D., Clements, J., i Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res *39*, W29-37.

Fukuchi, S., Sakamoto, S., Nobe, Y., Murakami, S.D., Amemiya, T., Hosoda, K., Koike, R., Hiroaki, H., i Ota, M. (2012). IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. Nucleic Acids Res *40*, D507-511.

Gajda, M.J., Tuszynska, I., Kaczor, M., Bakulina, A.Y., i Bujnicki, J.M. (2010). FILTREST3D: discrimination of structural models using restraints from experimental data. Bioinformatics *26*, 2986-2987.

George, R.A., i Heringa, J. (2002). Protein domain identification and improved sequence similarity searching using PSI-BLAST. Proteins *48*, 672-681.

George, R.A., Lin, K., i Heringa, J. (2005). Scooby-domain: prediction of globular domains in protein sequence. Nucleic Acids Res *33*, W160-163.

Greene, L.H., Lewis, T.E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A.*, et al.* (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res *35*, D291-297.

Griffin, M.D., i Gerrard, J.A. (2012). The relationship between oligomeric state and protein function. Adv Exp Med Biol *747*, 74-90.

Grigoryan, G., i Keating, A.E. (2008). Structural specificity in coiled-coil interactions. Curr Opin Struct Biol *18*, 477-483.

Gruber, M., Soding, J., i Lupas, A.N. (2005). REPPER--repeats and their periodicities in fibrous proteins. Nucleic Acids Res *33*, W239-243.

Guex, N., i Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis *18*, 2714-2723.

Guhaniyogi, J., i Brewer, G. (2001). Regulation of mRNA stability in mammalian cells. Gene *265*, 11-23.

Hao, B., Zhao, G., Kang, P.T., Soares, J.A., Ferguson, T.K., Gallucci, J., Krzycki, J.A., i Chan, M.K. (2004). Reactivity and chemical synthesis of L-pyrrolysine- the 22(nd) genetically encoded amino acid. Chem Biol *11*, 1317-1324.

Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M., i Iakoucheva, L.M. (2006a). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS computational biology *2*, e100.

Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M., i Iakoucheva, L.M. (2006b). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol *2*, e100.

Hirose, S., Shimizu, K., Kanai, S., Kuroda, Y., i Noguchi, T. (2007). POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. Bioinformatics *23*, 2046-2053.

Hofmann, K., i Stoffel, W. (1993). TMBASE - A database of membrane spanning proteins segments. Biol Chem Hoppe Seyler *347*, 166.

Holm, L., i Rosenstrom, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res *38*, W545-549.

Huang, H.L., Lin, I.C., Liou, Y.F., Tsai, C.T., Hsu, K.T., Huang, W.L., Ho, S.J., i Ho, S.Y. (2011). Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties. BMC Bioinformatics *12 Suppl 1*, S47.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S.*, et al.* (2012). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res *40*, D306-312.

Hwang, S., Gou, Z., i Kuznetsov, I.B. (2007). DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics *23*, 634-636.

Ishida, T., i Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res *35*, W460-464.

Ji, A., Johnson, M.T., Walsh, E.J., McGee, J., i Armstrong, D.L. (2013). Discrimination of individual tigers (Panthera tigris) from long distance roars. J Acoust Soc Am *133*, 1762-1769.

Jin, Y., i Dunbrack, R.L., Jr. (2005). Assessment of disorder predictions in CASP6. Proteins *61 Suppl 7*, 167-175.

Johansson, L., Gafvelin, G., i Arner, E.S. (2005). Selenocysteine in proteins-properties and biotechnological use. Biochim Biophys Acta *1726*, 1-13.

Jones, D.T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol *287*, 797-815.

Jones, D.T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics *23*, 538-544.

Jones, D.T., Taylor, W.R., i Thornton, J.M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry *33*, 3038-3049.

Jones, S., Daley, D.T., Luscombe, N.M., Berman, H.M., i Thornton, J.M. (2001). Protein-RNA interactions: a structural analysis. Nucleic Acids Res *29*, 943-954.

Jones, S., Stewart, M., Michie, A., Swindells, M.B., Orengo, C., i Thornton, J.M. (1998). Domain assignment for protein structures using a consensus approach: characterization and analysis. Protein science : a publication of the Protein Society *7*, 233-242.

Jones, S., van Heyningen, P., Berman, H.M., i Thornton, J.M. (1999). Protein-DNA interactions: A structural analysis. J Mol Biol *287*, 877-896.

Kabsch, W., i Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577-2637.

Kall, L., Krogh, A., i Sonnhammer, E.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res *35*, W429-432.

Karplus, K. (2009). SAM-T08, HMM-based protein structure prediction. Nucleic Acids Res *37*, W492-497.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M., i Hughey, R. (2003). Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins *53 Suppl 6*, 491-496.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., i Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Res *36*, D202-205.

Keerthi, S.S., i Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput *15*, 1667-1689.

Kim, P.M., Sboner, A., Xia, Y., i Gerstein, M. (2008). The role of disorder in interaction networks: a structural analysis. Mol Syst Biol *4*, 179.

Kirschner, A., i Frishman, D. (2008). Prediction of beta-turns and beta-turn types by a novel bidirectional Elman-type recurrent neural network with multiple output layers (MOLEBRNN).

Gene *422*, 22-29.

Kloczkowski, A., Ting, K.L., Jernigan, R.L., i Garnier, J. (2002). Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins *49*, 154-166.

Korneta, I., i Bujnicki, J.M. (2012). Intrinsic disorder in the human spliceosomal proteome. PLoS computational biology *8*, e1002641.

Korneta, I., Magnus, M., i Bujnicki, J.M. (2012). Structural bioinformatics of the human spliceosomal proteome. Nucleic acids research *40*, 7046-7065.

Kosinski, J., Gajda, M.J., Cymerman, I.A., Kurowski, M.A., Pawlowski, M., Boniecki, M., Obarska, A., Papaj, G., Sroczynska-Obuchowicz, P., Tkaczuk, K.L.*, et al.* (2005). FRankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6. Proteins *61 Suppl 7*, 106-113.

Kozlowski, L.P., i Bujnicki, J.M. (2012). MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinformatics *13*, 111.

Krogh, A., Larsson, B., von Heijne, G., i Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol *305*, 567-580.

Kumar, M., Gromiha, M.M., i Raghava, G.P. (2007). Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins.

Kurowski, M.A., i Bujnicki, J.M. (2003). GeneSilico protein structure prediction meta-server. Nucleic Acids Res *31*, 3305-3307.

Kyte, J., i Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J Mol Biol *157*, 105-132.

Letunic, I., Doerks, T., i Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res *40*, D302-305.

Levinthal, C. (1968). Are there pathways for protein folding? JChemPhys *65*, 44-45.

Levitt, M., i Chothia, C. (1976). Structural patterns in globular proteins. Nature *261*, 552-558.

Li, W., i Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Lin, H.H., i Tseng, L.Y. (2010). DBCP: a web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding state of cysteines. Nucleic Acids Res *38*, W503-507.

Lin HT, L.C. (2003). A study on sigmoid kernels for SVM and the training of non-PSD kernels

by SMO-type methods. Department of Computer Science, National Taiwan University.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., i Russell, R.B. (2003a). Protein disorder prediction: implications for structural proteomics. Structure *11*, 1453-1459.

Linding, R., Russell, R.B., Neduva, V., i Gibson, T.J. (2003b). GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res *31*, 3701-3708.

Lu, L., Niu, B., Zhao, J., Liu, L., Lu, W.C., Liu, X.J., Li, Y.X., i Cai, Y.D. (2009). GalNAc-transferase specificity prediction based on feature selection method. Peptides *30*, 359-364.

Lupas, A., Van Dyke, M., i Stock, J. (1991). Predicting coiled coils from protein sequences. Science *252*, 1162-1164.

Luscombe, N.M., i Thornton, J.M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol *320*, 991-1009.

Madera, M. (2008). Profile Comparer: a program for scoring and aligning profile hidden Markov models. Bioinformatics *24*, 2630-2631.

Magrane, M., i Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) *2011*, bar009.

Mandel, C.R., Bai, Y., i Tong, L. (2008). Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci *65*, 1099-1122.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R.*, et al.* (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res *39*, D225-229.

Margelevicius, M., i Venclovas, C. (2010). Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. BMC Bioinformatics *11*, 89.

Marsden, R.L., McGuffin, L.J., i Jones, D.T. (2002). Rapid protein domain assignment from amino acid sequence using predicted secondary structure. Protein Sci *11*, 2814-2824.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et biophysica acta *405*, 442-451.

McGuffin, L.J., Bryson, K., i Jones, D.T. (2000). The PSIPRED protein structure prediction server. Bioinformatics *16*, 404-405.

Millevoi, S., i Vagner, S. (2010). Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. Nucleic acids research *38*, 2757-2774.

Miyazawa, S., i Jernigan, R.L. (1999). Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins *34*, 49-68.

Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A., i Kryshtafovych, A. (2011). Evaluation of disorder predictions in CASP9. Proteins *79 Suppl 10*, 107-118.

Montgomerie, S., Cruz, J.A., Shrivastava, S., Arndt, D., Berjanskii, M., i Wishart, D.S. (2008). PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. Nucleic Acids Res *36*, W202-209.

Murakami, Y., Spriggs, R.V., Nakamura, H., i Jones, S. (2010). PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences. Nucleic Acids Res *38*, W412-416.

Murzin, A.G., Brenner, S.E., Hubbard, T., i Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol *247*, 536-540.

Nissen, S. (2003). Implementation of a Fast Artificial Neural Network Library (fann).

Noivirt-Brik, O., Prilusky, J., i Sussman, J.L. (2009). Assessment of disorder predictions in CASP8. Proteins *77 Suppl 9*, 210-216.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., i Dunker, A.K. (2005). Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins *61 Suppl 7*, 176-182.

Ofran, Y., Mysore, V., i Rost, B. (2007). Prediction of DNA-binding residues from sequence. Bioinformatics *23*, i347-353.

Ouali, M., i King, R.D. (2000). Cascaded multiple classifiers for secondary structure prediction. Protein Sci *9*, 1162-1176.

Pauling, L., i Corey, R.B. (1951). Configuration of polypeptide chains. Nature *168*, 550-551.

Pawlowski, M., i Bujnicki, J.M. (2012). The utility of comparative models and the local model quality for protein crystal structure determination by Molecular Replacement. BMC Bioinformatics *13*, 289.

Pawlowski, M., Gajda, M.J., Matlak, R., i Bujnicki, J.M. (2008). MetaMQAP: a meta-server for the quality assessment of protein models BMC Bioinformatics *9*, 403.

Pearl, F.M., Bennett, C.F., Bray, J.E., Harrison, A.P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., i Orengo, C.A. (2003). The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Res *31*, 452-455.

Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M., i Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol *9*, 51.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., i Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis.

J Comput Chem *25*, 1605-1612.

Pollastri, G., Baldi, P., Fariselli, P., i Casadio, R. (2002). Prediction of coordination number and relative solvent accessibility in proteins. Proteins *47*, 142-153.

Pollastri, G., i McLysaght, A. (2005). Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics *21*, 1719-1720.

Porter, L.L., i Rose, G.D. (2012). A thermodynamic definition of protein domains. Proceedings of the National Academy of Sciences of the United States of America *109*, 9420-9425.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., i Flannery, B.P. (2007). Numerical Recipes 3rd Edition: The Art of Scientific Computing (Cambridge University Press).

Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C., i Ouzounis, C.A. (2000). CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. Bioinformatics *16*, 915-922.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J.*, et al.* (2012). The Pfam protein families database. Nucleic Acids Res *40*, D290-301.

Puton, T., Kozlowski, L., Tuszynska, I., Rother, K., i Bujnicki, J.M. (2012). Computational methods for prediction of protein-RNA interactions. J Struct Biol *179*, 261-268.

Ramachandran, G.N., Ramakrishnan, C., i Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. Journal of molecular biology *7*, 95-99.

Redecke, L., Nass, K., Deponte, D.P., White, T.A., Rehders, D., Barty, A., Stellato, F., Liang, M., Barends, T.R., Boutet, S.*, et al.* (2012). Natively Inhibited Trypanosoma brucei Cathepsin B Structure Determined by Using an X-ray Laser. Science.

Remmert, M., Biegert, A., Hauser, A., i Soding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods *9*, 173-175.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., i Muller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

Rohl, C.A., Strauss, C.E., Misura, K.M., i Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol *383*, 66-93.

Rosner, B., Glynn, R.J., i Lee, M.L. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. Biometrics *62*, 185-192.

Rost, B., i Liu, J. (2003). The PredictProtein server. Nucleic Acids Res *31*, 3300-3304.

Rost, B., i Sander, C. (1993). Prediction of protein secondary structure at better than 70%

accuracy. J Mol Biol *232*, 584-599.

Rychlewski, L., Jaroszewski, L., Li, W., i Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci *9*, 232-241.

Sadreyev, R.I., Tang, M., Kim, B.H., i Grishin, N.V. (2007). COMPASS server for remote homology inference. Nucleic Acids Res *35*, W653-658.

Salamov, A.A., i Solovyev, V.V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J Mol Biol *247*, 11-15.

Sali, A., Potterton, L., Yuan, F., van, V.H., i Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. Proteins *23*, 318-326.

Savage, C.R., Jr., Hash, J.H., i Cohen, S. (1973). Epidermal growth factor. Location of disulfide bonds. J Biol Chem *248*, 7669-7672.

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., i Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. PLoS ONE *4*, e4433.

Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M., i Rost, B. (2011). Protein disorder--a breakthrough invention of evolution? Curr Opin Struct Biol *21*, 412-418.

Shameer, K., Pugalenthi, G., Kandaswamy, K.K., i Sowdhamini, R. (2011a). 3dswap-pred: prediction of 3D domain swapping from protein sequence using Random Forest approach. Protein Pept Lett *18*, 1010-1020.

Shameer, K., Shingate, P.N., Manjunath, S.C., Karthika, M., Pugalenthi, G., i Sowdhamini, R. (2011b). 3DSwap: curated knowledgebase of proteins involved in 3D domain swapping. Database : the journal of biological databases and curation *2011*, bar042.

Shannon, C.E. (1948). A Mathematical Theory of Communication. At&T Tech J *27*, 379-423.

Shen, H.B., i Chou, K.C. (2007). EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun *364*, 53-59.

Shi, Y., Di Giammartino, D.C., Taylor, D., Sarkeshik, A., Rice, W.J., Yates, J.R., 3rd, Frank, J., i Manley, J.L. (2009). Molecular architecture of the human pre-mRNA 3' processing complex. Mol Cell *33*, 365-376.

Shimizu, K., Hirose, S., i Noguchi, T. (2007). POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. Bioinformatics *23*, 2337-2338.

Siddiqui, A.S., i Barton, G.J. (1995). Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. Protein Sci *4*, 872-884.

Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., i

Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res *38*, D161-166.

Sim, J., Kim, S.Y., i Lee, J. (2005). PPRODO: Prediction of protein domain boundaries using neural networks. Proteins.

Smith, T.F., i Waterman, M.S. (1981). Identification of common molecular subsequences. JMolBiol *147*, 195-197.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics *21*, 951-960.

Sonnhammer, E.L., von Heijne, G., i Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol *6*, 175-182.

Su, C.T., Chen, C.Y., i Hsu, C.M. (2007). iPDA: integrated protein disorder analyzer. Nucleic Acids Res *35*, W465-472.

Su, C.T., Chen, C.Y., i Ou, Y.Y. (2006). Protein disorder prediction by condensed PSSM considering propensity for order or disorder. BMC Bioinformatics *7*, 319.

Suyama, M., i Ohara, O. (2003). DomCut: prediction of inter-domain linker regions in amino acid sequences. Bioinformatics *19*, 673-674.

Swindells, M.B. (1995). A procedure for detecting structural domains in proteins. Protein Sci *4*, 103-112.

Terribilini, M., Sander, J.D., Lee, J.H., Zaback, P., Jernigan, R.L., Honavar, V., i Dobbs, D. (2007). RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. Nucleic Acids Res *35*, W578-584.

Tjong, H., Qin, S., i Zhou, H.X. (2007). PI2PE: protein interface/interior prediction engine. Nucleic Acids Res *35*, W357-362.

Tompa, P. (2010). Intrinsically Disordered Proteins (Chapman & Hall).

Tung, C.H., i Yang, J.M. (2007). fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies. Nucleic Acids Res *35*, W438-443.

Tusnady, G.E., i Simon, I. (2001). The HMMTOP transmembrane topology prediction server. Bioinformatics *17*, 849-850.

Uversky, V.N., Oldfield, C.J., i Dunker, A.K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys *37*, 215-246.

Viklund, H., i Elofsson, A. (2004). Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci *13*, 1908-1917.

Viklund, H., i Elofsson, A. (2008). OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics *24*, 1662-1668.

Vucetic, S., Brown, C.J., Dunker, A.K., i Obradovic, Z. (2003). Flavors of protein disorder. Proteins *52*, 573-584.

Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J., Sikes, J.G.*, et al.* (2005). DisProt: a database of protein disorder. Bioinformatics *21*, 137-140.

Vullo, A., Bortolami, O., Pollastri, G., i Tosatto, S.C. (2006). Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. Nucleic Acids Res *34*, W164-168.

Wagner, M., Adamczak, R., Porollo, A., i Meller, J. (2005). Linear regression models for solvent accessibility prediction in proteins. J Comput Biol *12*, 355-369.

Wallner, B., i Elofsson, A. (2003). Can correct protein models be identified? Protein Sci *12*, 1073-1086.

Wallner, B., i Elofsson, A. (2005). Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics *21*, 4248-4254.

Walsh, I., Martin, A.J., Di Domenico, T., i Tosatto, S.C. (2012). ESpritz: accurate and fast prediction of protein disorder. Bioinformatics *28*, 503-509.

Walsh, I., Martin, A.J., Mooney, C., Rubagotti, E., Vullo, A., i Pollastri, G. (2009). Ab initio and homology based prediction of protein domains by recursive neural networks. BMC Bioinformatics *10*, 195.

Wang, G., i Dunbrack, R.L., Jr. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589-1591.

Wang, L., i Brown, S.J. (2006). BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res *34*, W243-248.

Wang, L., Huang, C., Yang, M.Q., i Yang, J.Y. (2010). BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol *4 Suppl 1*, S3.

Wang, L., Yang, M.Q., i Yang, J.Y. (2009). Prediction of DNA-binding residues from protein sequence information using random forests. BMC Genomics *10 Suppl 1*, S1.

Wang, Z., Zhao, F., Peng, J., i Xu, J. (2011). Protein 8-class secondary structure prediction using conditional neural fields. Proteomics *11*, 3786-3792.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., i Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol *337*,

635-645.

Wedemeyer, W.J., Welker, E., Narayan, M., i Scheraga, H.A. (2000). Disulfide bonds and protein folding. Biochemistry *39*, 7032.

Wheelan, S.J., Marchler-Bauer, A., i Bryant, S.H. (2000). Domain size distributions can predict domain boundaries. Bioinformatics *16*, 613-618.

Wolf, E., Kim, P.S., i Berger, B. (1997). MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci *6*, 1179-1189.

Wootton, J.C. (1994). Sequences with "unusual" amino acid composition. Curr Opin Struct Biol *4*, 413-421.

Wu, J., Liu, H., Duan, X., Ding, Y., Wu, H., Bai, Y., i Sun, X. (2009). Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics *25*, 30-35.

Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K., i Uversky, V.N. (2010). PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochimica et biophysica acta *1804*, 996-1010.

Yang, Q., i Doublie, S. (2011). Structural biology of poly(A) site definition. Wiley Interdiscip Rev RNA *2*, 732-747.

Yang, Z.R., Thomson, R., McNeil, P., i Esnouf, R.M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics *21*, 3369-3376.

Zamyatnin, A.A. (1972). Protein volume in solution. Prog Biophys Mol Biol *24*, 107-123.

Zhang, T., Faraggi, E., Xue, B., Dunker, A.K., Uversky, V.N., i Zhou, Y. (2012). SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J Biomol Struct Dyn *29*, 799-813.

Zhang, Y. (2012). http://zhanglab.ccmb.med.umich.edu/PSSpred

Zhou, H., i Zhou, Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins *58*, 321-328.

10. Dorobek naukowy

Prace opublikowane wchodzące w skład niniejszej pracy doktorskiej:

1. **Kozlowski LP**, Bujnicki JM MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. **BMC Bioinformatics** 2012, 13:111.

2. **Kozłowski, Ł.**, Orłowski, J., Bujnicki, J.M., "Structure prediction of alternatively spliced proteins" in "RNA splicing: The complete guide" Editors: Stefan Stamm, Christopher Smith, Reinhard Luehrmann, Wiley-Blackwell, 2011.

Inne prace, których współautorem jest autor rozprawy:

1. Puton T., **Kozlowski LP**, Rother K., Bujnicki J.M. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. **Nucleic Acids Res.** 2013, doi: 10.1093/nar/gkt101

2. Nakagome S, Mano S, **Kozlowski L**, Bujnicki JM, Shibata H, Fukumaki Y, Kidd JR, Kidd KK, Kawamura S, Oota H Crohn's disease risk alleles on the NOD2 locus have been maintained by natural selection on standing variation. **Mol Biol Evol.** 2012 Jun;29(6):1569-8.

3. Al-Haggar M, Madej-Pilarczyk A, **Kozlowski L**, Bujnicki JM, Yahia S, Abdel-Hadi D, Hamed S, Puzianowska-Kuznicka M A novel homozygous p.Arg527Leu LMNA mutation in two unrelated Egyptian families causes overlapping mandibuloacral dysplasia and progeria syndrome. **Eur J Human Genet** 2012, 2012 Nov;20(11):1134-40.

4. Puton T., **Kozlowski L.**, Tuszynska I., Rother K., Bujnicki J.M. Computational methods for prediction of protein-RNA interactions J Struct Biol 2011 **J Struct Biol**. 2012 Sep;179(3):261-8.

5. Majorek, K., **Kozłowski, Ł.**, Jakalski, M., Bujnicki, J.M. rozdział "First steps of protein structure prediction" w książce "Prediction of Protein Structures, Functions and Interactions" pod redakcją: Janusz M. Bujnicki. Wiley & Sons 2008.

6. Pałyga J., **Kozłowski, Ł.** Structure and function of molecular chaperone HSP90. Sowriemiennyj Naucznyj Wiestnik Ser. Biologija Chimija 2007, 15(23).