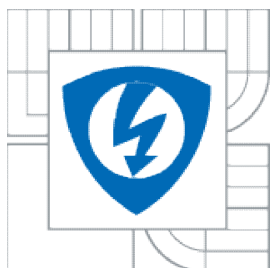




VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH
TECHNOLGIÍ
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF BIOMEDICAL ENGINEERING

PREDIKCE AKTIVNÍCH MÍST V PROTEINECH

PROTEIN HOT SPOTS PREDICTION

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

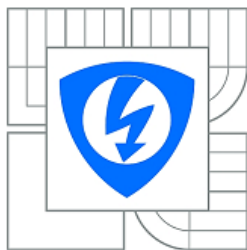
AUTOR PRÁCE
AUTHOR

Bc. JAN KAŠPÁREK

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. DENISA MADĚRÁNKOVÁ

BRNO 2013



VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

Ústav biomedicínského inženýrství

Diplomová práce

magisterský navazující studijní obor
Biomedicínské inženýrství a bioinformatika

Student: Bc. Jan Kašpárek

ID: 119718

Ročník: 2

Akademický rok: 2012/2013

NÁZEV TÉMATU:

Predikce aktivních míst v proteinech

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši používaných metod pro predikci aktivních míst (hot spots) v proteinech. Zaměřte se především na metody založené na EIIP hodnotách aminokyselin. 2) Navrhněte pseudokód a vývojový diagram pro vybranou metodu predikce aktivních míst založenou na EIIP. 3) Vybranou metodu implementujte v libovolném programovém prostředí. 4) Realizovaný algoritmus validujte na souboru dat a výsledky zhodnoťte srovnáním s publikovanými daty, případně proveďte srovnání získaných výsledků s výsledky z volně dostupných nástrojů pro predikci aktivních míst v proteinech.

DOPORUČENÁ LITERATURA:

[1] FERNANDEZ-RECIO, J. Prediction of protein binding sites and hot spots. WIREs Computational Molecular Science, 2011, roč. 1, s. 680-698.

[2] SAHU, S., PANDA S. Efficient Localization of Hot Spots in Proteins Using a Novel S-Transform Based Filtering Approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, roč. 8, no. 5, s. 1235-1246.

Termín zadání: 11.2.2013

Termín odevzdání: 24.5.2013

Vedoucí práce: Ing. Denisa Maděránková

Konzultanti diplomové práce:

prof. Ing. Ivo Provazník, Ph.D.

Předseda oborové rady

UPOZORNĚNÍ:

Autor diplomové práce nesmí při vytváření diplomové práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

Abstrakt

Znalost aktivních míst proteinů a schopnost určit jejich polohu pouze z informace o primární struktuře daného proteinu je dlouhodobým cílem snažení mnoha vědců po celém světě. Tato práce osvětluje význam aktivních míst jako takových a shrnuje pokroky dosažené na poli jejich predikce.

Kromě toho je zde představen predikční algoritmus pracující pouze s informacemi z primární struktury proteinů. Jeho základem jsou techniky zpracování signálů. Pro převod na numerický signál je využito veličiny EIIP a další zpracování probíhá pomocí S-transformace. Algoritmus dosahuje senzitivity větší než 60 %, jeho pozitivní prediktivní hodnota přesahuje 50 % a oproti jiným současným metodám vyniká zejména svou rychlostí a jednoduchostí.

Klíčová slova

Predikce aktivních míst, EIIP, S-transformace, digitální filtrace, sekvence proteinu, RRM

Abstract

Knowledge of protein hot spots and the ability to successfully predict them while using only primary protein structure has been a worldwide scientific goal for several decades. This thesis describes the importance of hot spots and sums up advances achieved in this field of study so far.

Besides that we introduce hot spot prediction algorithm using only a primary protein structure, based primarily on signal processing techniques. To convert protein sequence to numerical signal we use the EIIP attribute, while further processing is carried out via means of S-transform. The algorithm achieves sensitivity of more than 60 %, positive predictive value exceeds 50 % and the main advantage over competitive algorithms is its simplicity and low computational requirements.

Key words

Hot spots prediction, active sites, EIIP, S-transform, digital filters, protein sequence, RRM

KAŠPÁREK, J. *Predikce aktivních míst v proteinech*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2013. 64 s. Vedoucí diplomové práce Ing. Denisa Maděránková.

Prohlášení

Prohlašuji, že svou diplomovou práci na téma Predikce aktivních míst v proteinech jsem vypracoval samostatně pod vedením vedoucího semestrálního projektu a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením tohoto projektu jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009Sb.

V Brně dne 21. 5. 2013

.....
podpis autora

Poděkování

Děkuji vedoucí diplomové práce Ing. Denise Maděránkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady, které mi poskytla při zpracování mé diplomové práce.

V Brně dne 21. 5. 2013

.....
podpis autora

Obsah

Úvod.....	11
1 Proteiny a jejich komplexy.....	12
1.1 Vazebná místa.....	14
1.2 Aktivní místa.....	14
2 Mutageneze alaninovým vyhlazením.....	17
2.1 Změna volné vazebné energie.....	18
2.2 Změna volného povrchu molekuly.....	18
2.3 Množství dodnes testovaných reziduí.....	19
3 Výpočetní metody predikce aktivních míst.....	20
3.1 Metody založené na struktuře komplexu.....	21
3.1.1 MAPPIS.....	21
3.1.2 Hotpoint.....	21
3.1.3 ROBETTA.....	21
3.2 Metody založené na struktuře volných proteinů.....	22
3.2.1 pyDockNIP.....	22
3.3 Metody založené na sekvenci proteinu.....	22
3.3.1 ISIS.....	22
3.3.2 Využití digitální filtrace.....	23
3.3.3 Využití S-transformace.....	25
4 Navrhovaný algoritmus.....	27
4.1 Obecný popis algoritmu, metoda A.....	27
4.1.1 Vstupní data.....	27
4.1.2 Převod na hodnoty EIIP.....	28
4.1.3 Zarovnání.....	28
4.1.4 Výpočet konsenzuálního spektra.....	29
4.1.5 S-transformace.....	29
4.1.6 Násobení ST spektra a konsenzuálního spektra.....	31

4.1.7	Inverzní S-transformace a zpracování jejího výstupu.....	31
4.1.8	Prahování	31
4.2	Metoda B.....	32
4.2.1	Charakteristická frekvence.....	33
4.3	Metoda C.....	33
4.4	Metoda D.....	35
5	Pseudokód a vývojové diagramy	37
6	Dosažené výsledky.....	42
6.1	Referenční data	42
6.2	Metoda A.....	43
6.3	Metoda B.....	44
6.4	Metoda C.....	46
6.5	Metoda D.....	49
7	Diskuze.....	51
7.1	Srovnání navrhovaných metod.....	51
7.1.1	Nastavení SEP.....	51
7.1.2	Nastavení PPV_{max}	52
7.1.3	SEP versus PPV_{max}	52
7.2	Použití serveru Hotpoint jako reference.....	53
7.3	Vliv volby příbuzných proteinů	54
7.3.1	Zachování charakteristické frekvence.....	54
7.3.2	Změna charakteristické frekvence	55
7.4	Srovnání s výsledky jiných skupin.....	56
8	Popis přiloženého softwaru	57
8.1	Vstupní data	57
8.2	Volání funkce.....	57
8.3	Výstup analýzy.....	57
	Závěr	58
	Seznam literatury	60
	Seznam zkratk	64

Seznam obrázků

Obrázek 1. Obecný vzorec aminokyseliny	12
Obrázek 2. Kvartérní struktura hemoglobinu	13
Obrázek 3. Schematické znázornění velikosti vazebných a aktivních míst na příkladu vazby lidského růstového hormonu na jeho extracelulární receptor. Převzato z [15].....	15
Obrázek 4. Srovnání alaninu (vlevo) a izoleucinu (vpravo).....	17
Obrázek 5. Korelace $\Delta\Delta G$ a $\Delta\Delta SASA$. Převzato z [18].....	19
Obrázek 6. Srovnání STFT, CWT a ST; (a) Signál, (b) STFT spektrum, (c) CWT spektrum, (d) ST spektrum. Převzato z [25].....	26
Obrázek 7. Metoda A – blokové schéma	27
Obrázek 8. Signál lidského interleukinu 4 po předzpracování	28
Obrázek 9. Konsenzuální spektrum interleukinu 4	29
Obrázek 10. ST spektrum lidského interleukinu 4	30
Obrázek 11. ST spektrum lidského interleukinu 4 po pronásobení s konsenzuálním spektrem	30
Obrázek 12. Vlevo nahoře - výstup IST; vpravo nahoře - výstup IST po umocnění na druhou (modře) a po vyhlazení (červeně); dole - vyhlazený signál po opětovném umocnění (modře), aktivní místa dle ASEdb (červené křížky), residua v ASEdb neklasifikovaná jako aktivní místa (černé křížky), detekční práh našeho algoritmu (černá linie)	32
Obrázek 13. Metoda B – blokové schéma	32
Obrázek 14. Metoda B - průběh charakteristické frekvence interleukinu 4 (modře), práh (černě), aktivní místa podle ASEdb (červené křížky), rezidua podle ASEdb neklasifikovaná jako aktivní místa (černé křížky).....	33
Obrázek 15. Metoda C – blokové schéma	34
Obrázek 16. Metoda C – detekce; vyhlazený a umocněný výstup IST (modře), práh (černě), aktivní místa dle ASEdb (červené křížky), rezidua dle ASEdb neklasifikovaná jako aktivní místa (černé křížky), rezidua algoritmem detekovaná jako aktivní místa (kroužky).....	34
Obrázek 17. Metoda D – blokové schéma.....	35
Obrázek 18. Metoda D – detekce; řez ST spektrem na charakteristické frekvenci (modře), práh (černě), aktivní místa podle ASEdb (červené křížky), rezidua podle ASEdb neklasifikovaná jako aktivní místa (černé křížky), rezidua detekovaná algoritmem jako aktivní místa (kroužky)	35
Obrázek 19. Vývojový diagram pro funkci NAJDI_AM	38
Obrázek 20. Vývojový diagram pro funkci HAMMING	38

Obrázek 21. Vývojový diagram pro funkci AA_NA_EIIP	39
Obrázek 22. Metoda A – závislost statistických ukazatelů na volbě prahu.....	43
Obrázek 23. Metoda B – závislost statistických ukazatelů na volbě prahu	45
Obrázek 24. Metoda C – závislost statistických ukazatelů na volbě prahu	46
Obrázek 25. Metoda C – závislost senzitivity na volbě prahu a šířky detekce.....	47
Obrázek 26. Metoda C – závislost specificity na volbě prahu a šířky detekce.....	47
Obrázek 27. Metoda C – závislost pozitivní prediktivní hodnoty na volbě prahu a šířky detekce	48
Obrázek 28. Metoda D – závislost senzitivity na volbě prahu a šířky detekce	49
Obrázek 29. Metoda D – závislost specificity na volbě prahu a šířky detekce	49
Obrázek 30. Metoda D – závislost pozitivní prediktivní hodnoty na volbě prahu a šířky detekce	50
Obrázek 31. Srovnání statistických ukazatelů	52
Obrázek 32. ROC křivka pro metody A a B.....	53

Seznam tabulek

Tabulka 1. Hodnoty EIIP pro jednotlivé aminokyseliny. Převzato z [30].....	23
Tabulka 2. Základní informace o referenčních proteinech	42
Tabulka 3. Vybrané proteiny - zdrojový organismus a příbuzné proteiny	42
Tabulka 4. Statistické údaje o účinnosti metody A	44
Tabulka 5. Metoda A – souhrnné ukazatele účinnosti	44
Tabulka 6. Statistické údaje o účinnosti metody B.....	45
Tabulka 7. Metoda B – souhrnné ukazatele účinnosti	46
Tabulka 8. Metoda C – souhrnné ukazatele účinnosti	48
Tabulka 9. Metoda D – souhrnné ukazatele účinnosti.....	50
Tabulka 10. Kategorizace navrhovaných metod.....	51
Tabulka 11. Srovnání přehledových ukazatelů navrhovaných metod	51
Tabulka 12. Srovnání souhrnných ukazatelů za použití referenčních dat ze serveru Hotpoint	53
Tabulka 13. Řetězce referenčních proteinů pro srovnání s Hotpointem.....	54
Tabulka 14. Změna účinnosti při odebrání P30367	55
Tabulka 15. Změna účinnosti při odebrání P30368.....	55
Tabulka 16. Srovnání navrhovaných metod s jinými metodami. Údaje částečně převzaty z [9], [25]	56

Úvod

Proteiny jsou stavebním kamenem těl živých organismů. Snad ještě důležitější je ale jejich úloha coby katalyzátorů, případně inhibitorů, chemických reakcí. Právě proteiny umožňují, aby v našich tělech probíhaly procesy, které by za běžných podmínek byly nemožné. Tato role proteinů je zprostředkována proteinovými komplexy, které se samy skládají z různého počtu samostatných proteinových řetězců. Teprve výsledná podoba těchto komplexů umožňuje jejich správnou funkci. Proto je žádoucí, abychom byli schopni určit, které proteinové řetězce jsou spolu schopny tyto komplexy tvořit, nebo jejich tvorbu dokonce řídit.

Důležitým poznatkem je zde právě znalost vazebných míst, díky kterým se proteinové jednotky slučují do komplexů. Vazebná místa jsou ale stále ještě příliš velká. Víme však, že více než 90 % vazebné energie má na svědomí jen několik málo jejich reziduí, ta jsou označována jako aktivní místa. Za posledních třicet let bylo aplikováno několik stěžejních myšlenek umožňujících tato místa lokalizovat. Laboratorní metody, jejichž spolehlivost dosud nebyla překonána, se však ukázaly jako časově náročné a drahé. Proto je zde velký tlak na vývoj výpočetních modelů predikujících aktivní místa ze samotné sekvence proteinu, což je informace, kterou jsme na rozdíl od jejich struktury schopni efektivně zjistit.

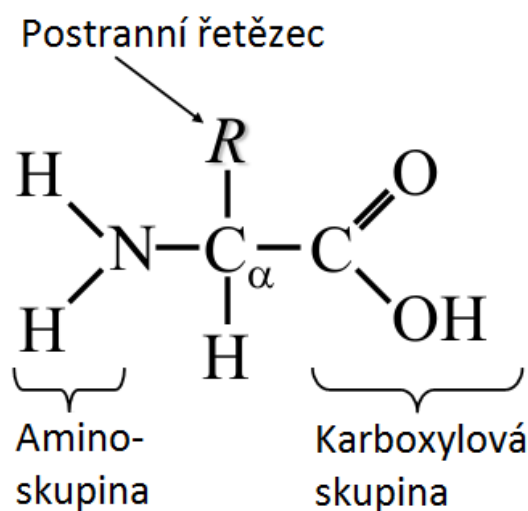
Znalost aktivních míst napomůže nejen k pochopení samotných metabolických drah a jejich vzájemných souvislostí, farmaceutický průmysl si od ní slibuje i možnost návrhu léčiv schopných tyto pochody efektivně ovlivnit a léčit tak choroby dnes považované za neléčitelné.

Tato práce obsahuje přehled metod používaných k predikci aktivních míst proteinů s použitím výpočetní techniky. Zaměřuji se přitom na metody používající jako vstup výhradně sekvenci aminokyselin tvořících daný protein. Kromě toho je zde představen jednoduchý predikční algoritmus založený na technikách zpracování signálů, S-transformaci a konverzi proteinové sekvence pomocí veličiny EIIP.

1 Proteiny a jejich komplexy

Proteiny jsou polymery organického původu, jejichž základními stavebními jednotkami jsou aminokyseliny. Aminokyselin v přírodě existuje nepřeberné množství, lidský genom jich ovšem kóduje dvacet. Při průměrné délce proteinu 644 aminokyselin [1] je tak množství kombinací ohromné. Samozřejmě ne všechny kombinace se v přírodě uskutečňují a ani z reálných proteinů nejsou všechny pro výzkum zajímavé.

Obecnou strukturu aminokyseliny popisuje obrázek 1. Aminoskupina se spolu s karboxylovou skupinou vyskytuje u všech aminokyselin beze změny. Obě skupiny se účastní peptidové vazby, díky níž se jedna aminokyselina váže na druhou a tvoří tak peptid, případně protein. Odlišný bývá postranní řetězec na obrázku reprezentovaný písmenem R. U nejjednodušší aminokyseliny glycinu je zastoupen pouze vodíkovým atomem, zatímco například u tryptofanu se zde přes vložený uhlíkový atom nachází navázaný celý purin. Právě podoba postranního řetězce udává vlastnosti aminokyseliny. Podle jejich vlastností je pak možno dělit je do skupin. Nejběžnější dělení zahrnuje aminokyseliny polární (hydrofilní) a nepolární (hydrofobní) [2]. Přehled všech proteinogenních aminokyselin kódovaných v lidském genomu je uveden v tabulce 1, jejich strukturu zde však až na výjimky neuvádím.



Obrázek 1. Obecný vzorec aminokyseliny

Pro ilustraci důležitosti, jakou proteiny v organismech mají, je vhodné popsat samotný proces vzniku proteinu. Syntéza proteinů probíhá v jednotlivých doménách organismů mírně odlišně, následující popis proto bude brát v úvahu děje odehrávající se v eukaryotické buňce, potažmo lidském organismu.

Prvním krokem syntézy je transkripce, během níž dochází k přepisu informace z jaderné DNA do mediátorové RNA (mRNA). Tento přepis je realizován enzymem RNA polymerázou, jenž je sám, stejně jako drtivá většina ostatních enzymů, proteinem.

Takto syntetizovaná mRNA podléhá splicingu v jaderné organelle zvané spliceozom. Splicing probíhá za účasti molekul snRNA (small nuclear RNA, tedy malá jaderná RNA) a většího množství dalších proteinů [3]. Upravená sekvence mRNA je dále transportována z jádra do cytoplasmy buňky, kde podléhá procesu translace. Při něm sekvence bází nukleové kyseliny hraje roli templátu, podle něž se na ribozomech syntetizuje již finální sekvence aminokyselin tvořící primární strukturu proteinu. Funkce proteinu však je kromě jeho primární struktury dána i jeho vyšší strukturou. Nově vzniklým proteinům proto často asistují chaperony, aby zajistily jejich správné sbalení do funkční struktury. Chaperony jsou opět proteiny.

Vidíme, že už při samotné syntéze proteinů, je téměř v každém kroku potřebná účast jiného, již syntetizovaného, proteinu. Tato skutečnost je výmluvným důkazem toho, jak důležitou roli tyto makromolekuly v organismech hrají. Ve skutečnosti je jejich účast nezbytná při drtivé většině buněčných pochodů.

Proteiny syntetizované procesem popsaným výše se dále často spojují do proteinových komplexů. Tvorba proteinového komplexu značí interakci proteinů. Interagovat může dvojice proteinů, což je nejčastější, známe ale i komplexy složené z více než sta podjednotek [4]. Interagují-li dva identické proteiny, vzniká homokomplex. Interakcí dvou rozdílných proteinů vzniká naopak heterokomplex, často méně stabilní [4]. Příkladem proteinového komplexu může být hemoglobin skládající se ze čtyř proteinových podjednotek (Obrázek 2), dvou α a dvou β jednotek. Imunoglobuliny jsou také tvořeny čtyřmi podjednotkami, dvěma lehkými a dvěma těžkými. Cytochrom C za určitých podmínek tvoří homodimer [5]. Příkladem zvláště velkých proteinů mohou být membránové proteiny [6].



Obrázek 2. Kvartérní struktura hemoglobinu

Komplexy jsou tvořeny při veškerých interakcích proteinů. Kromě komplexů tvořících funkční celek známe i komplexy typu enzym-inhibitor, kdy je komplex naopak neaktivní formou proteinu. Stejně tak navázáním proteinu na receptor opět vzniká komplex, což je v podstatě i případ vazby protilátka-antigen.

1.1 Vazebná místa

Mohlo by se zdát, že proteiny se mezi sebou mohou vázat libovolně, opak je však pravdou. Jejich vzájemné interakce jsou vysoce specifické. Jednotlivé proteinové jednotky na svém povrchu nesou vazebná místa, jejichž tvar a složení dovoluje navázání pouze určitých vyhovujících partnerů. Plocha vazebných míst se podle Jonese pohybuje od 368 \AA^2 do 4746 \AA^2 ($\text{\AA} = 0,1 \text{ nm}$) [4]. Hodnoty dále závisejí na konkrétním druhu interakce. Uvádím sice největší možný rozptyl, hodnoty u jednotlivých druhů vazeb ale nejsou o mnoho přesnější.

Vazebná místa jsou tedy relativně velká, Janin [7] uvádí, že u menších proteinů může jejich plocha dosahovat až dvaceti procent celkového povrchu, přičemž v jeho práci jsou zmiňovány podstatně menší plochy vazebných míst. Obě zmiňované práce však operují pouze s omezeným vzorkem proteinů, a to pokaždé jiným. Lze proto očekávat, že jejich výsledky tak mohou být zkresleny.

S přihlédnutím k těmto údajům není žádným překvapením, že vazebná místa jsou tvořena desítkami aminokyselin [8]. Přes svůj rozměr se často jedná o relativně ploché oblasti [9], přesto bylo vyvinuto několik metod predikujících vazebná místa založených právě na sterické kompatibilitě obou partnerů. Rajamani v naopak tvrdí, že na vazebných místech často nacházíme rezidua vystupující do prostoru, nebo naopak jamky, do kterých tyto řetězce mohou zapadnout [10]. Tento názor nelze vyvrátit, rozměry výstupků, případně výdutí, jsou ale v porovnání s rozlohou vazebných míst na první pohled zanedbatelné, i když funkčně bezpochyby zcela zásadní.

Druh interakce proteinů kromě velikosti jejich vazebných míst ovlivňuje i jejich složení. Zatímco u homodimerů jsou vlastnosti vazebných míst jasně rozdílné od zbytku povrchu proteinu, kde převažují polární aminokyseliny, u heterodimerů jednoznačně rozdílné vlastnosti pozorovat nemůžeme. [9]

1.2 Aktivní místa

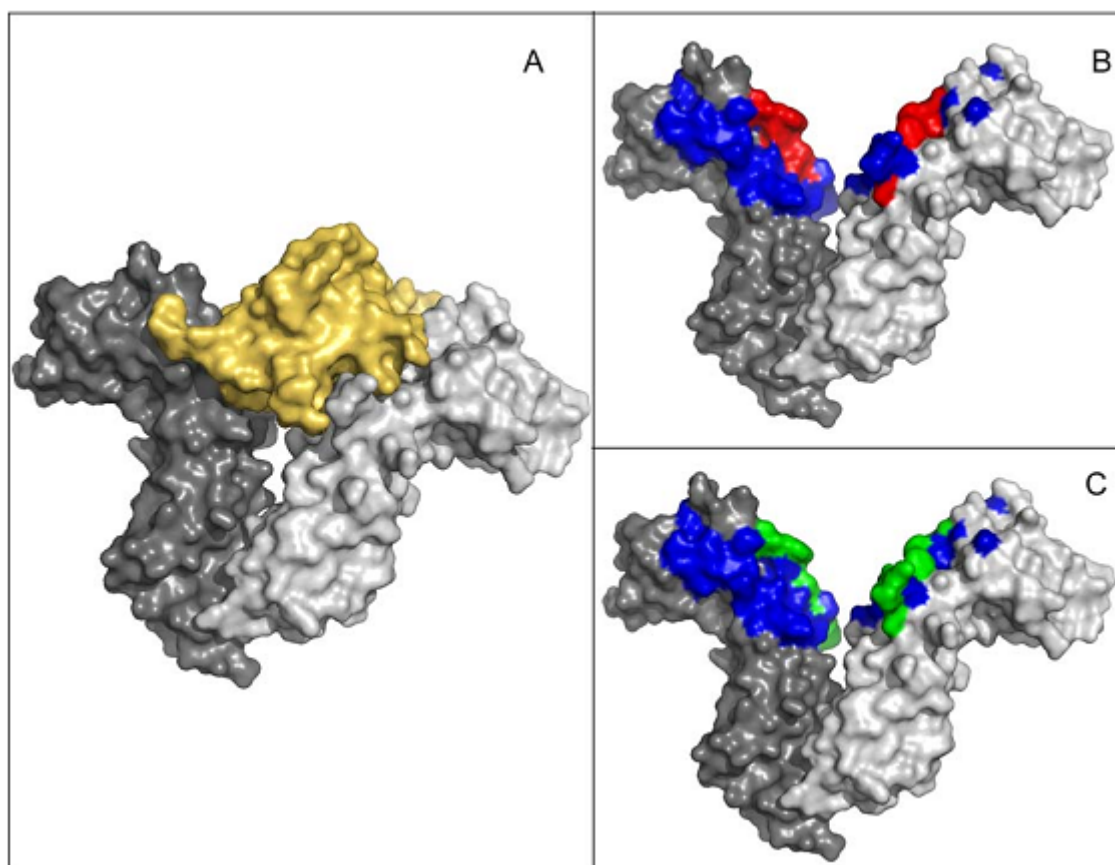
Ne všechny aminokyseliny vazebných míst jsou pro tvorbu komplexu stejně důležité. Experimentálními metodami bylo zjištěno, že pouze malé množství aminokyselin nacházejících se ve vazebných místech proteinů výrazně přispívá k celkové vazebné energii. Jejich počet je tak nízký, že se skutečně jedná jen o jednotlivá rezidua. Na povrchu proteinů se však často vyskytují v souvislých shlucích [11], rezidua sousedící z hlediska prostorového ovšem nemusejí sousedit při pohledu na sekvenci. Pojem aktivní místa potom bývá používán jak pro označení konkrétních reziduí, tak pro označení celých jejich shluků.

Právě tyto aminokyseliny jsou považovány za hlavní příčinu specifity proteinových interakcí [12]. V zahraniční literatuře bývají označovány jako *hot spots*, Rajamani je v [10] označuje jako „kotevní aminokyseliny“ (*anchor residues*), a to i přesto, že se v jeho práci vyskytuje i pojem *hot spot*. V dnešní literatuře je však

již sousloví *hot spot* zavedeným pojmem [13]. V českých publikacích je potom používán termín aktivní místa, nebo je přejat anglický výraz.

Aktivní místa jsou vysoce konzervativní [9], nacházíme je proto u různých organismů často identické, nebo jen s malými změnami. Tato skutečnost je využívána v metodách založených na evoluční informaci. Často je nacházíme poblíž středu vazebných míst, tuto skutečnost poprvé objevil Clackson [11] na lidském růstovém hormonu a později ji Bogan prokázal v obecném měřítku [14]. Ani výskyt aktivních míst na okrajích vazebných míst však není vyloučen a některé takové případy jsou i pozorovány [14].

Aktivní místa bývají obklopeny dalšími, energeticky méně důležitými, konzervativními aminokyselinami tvořícími ochranný hydrofobní prstenec nazývaný podle široce rozšířených těsnících kroužků *O-ring* [4]. Přitom se ukazuje, že vypuzení vody pryč od aktivních míst je důležitým předpokladem pro jejich vysoce energetickou vazbu v komplexu [14].



Obrázek 3. Schematické znázornění velikosti vazebných a aktivních míst na příkladu vazby lidského růstového hormonu na jeho extracelulární receptor. Převzato z [15]

Nejčtenější aminokyseliny představující aktivní místa proteinu jsou podle informací ze článku [14] tryptofan, arginin a tyrosin. Naopak nejméně zastoupenými jsou serin, leucin a threonin. Valin a cystein dokonce nebyly v aktivních

místech vůbec pozorovány. V případě cysteinu však je jeho nepřítomnost důsledkem obtížné mutace na alanin zmiňované i v [11] (viz alaninové skenování v další kapitole). Tyto údaje jsou v dobré shodě s [8], kde je použit poněkud odlišný přístup.

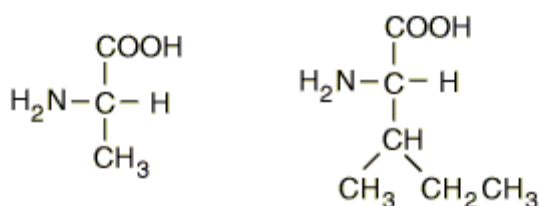
Aktivní místa svou malou velikostí a svým majoritním podílem vazebné energie představují ideální cíl pro ovlivnění proteinových interakcí nasazením specifických léčiv. Zvlášť zajímavým cílem jsou proteiny signálních drah, ty totiž představují uzlové body mnoha buněčných pochodů. Motivace pro jejich další výzkum tedy není zanedbatelná.

Dobrou ilustraci velikostí vazebných a aktivních míst ve srovnání s velikostí proteinů nabízí obrázek 3. V části A vidíme lidský růstový hormon (žlutá) navázaný na homodimer svého extracelulárního receptoru (každá podjednotka má jiný odstín šedé). V části B je zobrazen receptor bez navázaného hormonu s vyznačenými vazebnými místy (modře) a aktivními místy určenými pomocí mutagenese alaninovým vyhlazením (červeně). Stejně rozložení je i v části C, kde modrá opět reprezentuje vazebná místa, zelená potom aktivní místa detekovaná metodou ISIS (viz kapitola 3.3.1). Každý řetězec receptoru je složen z 201 reziduí. Z nich 31 leží uvnitř vazebného místa. Vazebné místo tedy tvoří něco přes 15 % všech reziduí receptoru [15].

2 Mutageneze alaninovým vyhlazením

Prvotní snaha o nalezení aktivních míst probíhala pochopitelně v laboratořích. Na rozdíl od výpočetních technik, z těch laboratorních se prakticky využívá jediný přístup. Je jím mutageneze alaninovým vyhlazením (alanine scanning mutagenesis, ASM), jindy též kvůli anglickému názvu označovaná jako mutageneze alaninovým skenováním, nebo jen alaninové skenování.

Základem metody je postupná mutace všech aminokyselin proteinu na alanin tak, aby v každém kroku byla zmutována právě jedna aminokyselina. Takto vytvoření mutanti poté podléhají interakci se svými protějšky. Přitom se měří dva parametry reakce, které jsou považovány za hlavním kritérium pro určování aktivních míst: změna volné vazebné energie a změna volného povrchu molekuly přístupného rozpouštědлу.



Obrázek 4. Srovnání alaninu (vlevo) a izoleucinu (vpravo)

Obrázek 4 ilustruje rozdíl mezi alaninem a jednou z větších aminokyselin, izoleucinem. Obě aminokyseliny jsou zde bohužel vykresleny v jiném uspořádání než na obrázku 1. Na tomto srovnání je vidět, že mutace rezidua na alanin prakticky eliminuje vliv jeho postranního řetězce. Při celkovém pohledu na protein tak dojde k vyhlazení jeho hlavního řetězce, odtud český název metody.

Při proteinech o délce v řádu stovek reziduí je proces mutageneze alaninovým vyhlazením, značně zdlouhavý a pracný. Pro každé jednotlivé reziduum je totiž třeba vyvinout speciální kmen bakterií (často se používá *Escherichia coli* [11]), který jej bude produkovat v dostatečném množství, aby parametry následné reakce byly měřitelné. Následně je potřeba všechny mutované varianty nechat zreagovat a výsledky analyzovat. Jedná se tak sice o relativně spolehlivou metodu, která je výpočetními metodami často brána jako referenční, avšak její časová a finanční náročnost zabraňují plošnému nasazení při mapování proteinů. ASM je dokonce tak náročné, že v praxi ani proteiny, které jsou mu podrobeny, nejsou analyzovány v celé své délce

(viz kapitola 6.1). Místo toho jsou mutována jen rezidua, která jsou pokládána za zajímavá, přesto jich může být více než sto [16].

2.1 Změna volné vazebné energie

Předpokladem je, že síla vazby bude mutací ovlivněna. Hodnoty naměřené při reakcích mutantů jsou porovnávány s hodnotami dosaženými nezmutovaným proteinem. Žádná nebo malá změna vazebné energie naznačuje, že zmutovaná aminokyselina se vazby neúčastní. Naopak velký úbytek energie znamená, že daná aminokyselina má na vazbě významný podíl a představuje tak aktivní místo. Pokles volné vazebné energie dán mutací jednoho rezidua na alanin může překročit i hodnotu 4,5 kcal/mol (1 kcal = 4,185 kJ) [11]. Obecně se však jako práh pro klasifikaci aktivního místa uvažuje už hodnota 1 kcal/mol, případně 2 kcal/mol [9].

Vědecká obec v tomto není jednotná a v některých pracích se objevují i jiné hodnoty [15]. Byly dokonce zaznamenány případy, kdy místo poklesu vazebné energie došlo po mutaci rezidua na alanin k jejímu zvýšení. Šlo přitom o ojedinělá rezidua nacházející se mimo souvislý shluk aktivního místa [11].

Protože slovní vyjádření může být matoucí, uvádím pro jistotu i vzorec

$$\Delta\Delta G = \Delta G_{WT} - \Delta G_{MUT}, \quad (1)$$

kde $\Delta\Delta G$ je změna volné vazebné energie, ΔG_{WT} je volná vazebná energie původního komplexu a ΔG_{MUT} je volná vazebná energie komplexu s mutovaným proteinem. Je tedy jasně vidět, že kladná hodnota $\Delta\Delta G$ představuje úbytek volné vazebné energie a tím pádem méně stabilní protein při mutaci daného rezidua na alanin.

2.2 Změna volného povrchu molekuly

Kromě volné vazebné energie proteinů v komplexu se uvažuje i další kritérium. Velikost vazebných míst proteinů byla již dříve určována pomocí veličiny v zahraniční literatuře označované jako *solvent accessible surface area* (SASA). Do češtiny lze přeložit jako „volný povrch přístupný rozpouštědлу,“ neexistuje však žádný obecně používaný český ekvivalent. V dalším textu budu proto používat zkratku anglického označení.

Základním předpokladem je zde to, že mutací rezidua bude ovlivněna těsnost, s jakou na sebe podjednotky komplexu přiléhají a tím pádem i velikost povrchu výsledného komplexu. Podobně jako u měření volné vazebné energie, zde se měří úbytek SASA vzniklého proteinového komplexu [7]. Sloučením dvou proteinů (nebo obecně jakýchkoli částic) dojde ke zmenšení velikosti plochy přístupné rozpouštědлу [17]. Tato změna je měřitelná. Opět se uvažuje, že mutací aktivního místa dojde ke zmenšení úbytku SASA (tj. plocha přístupná rozpouštědлу bude větší). Jako prahová hodnota zde bývá uvažováno 100 Å² [10].

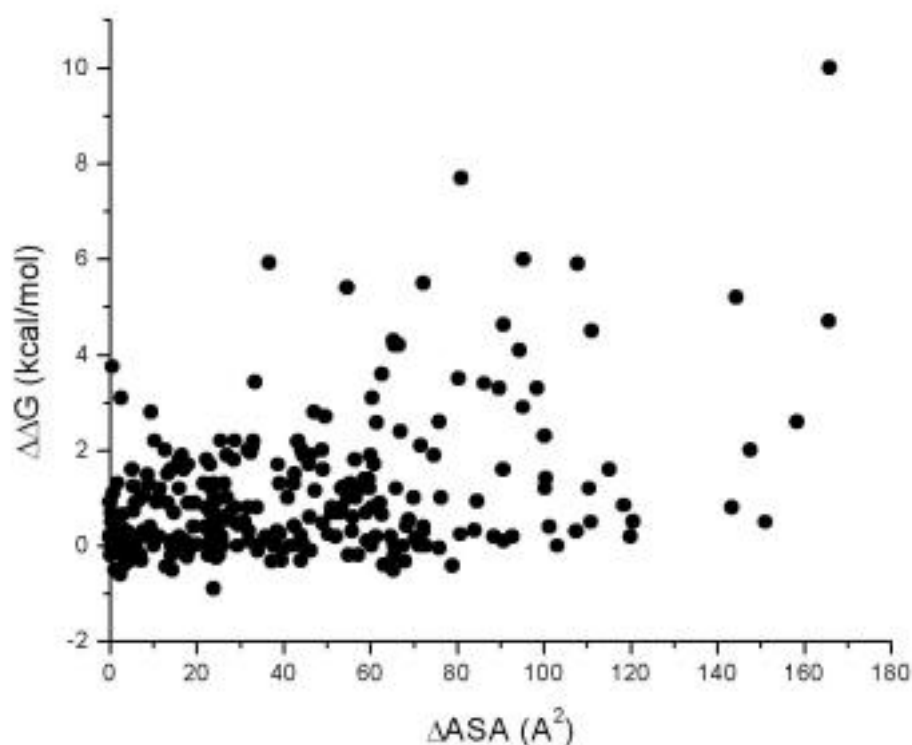
Protože slovní vyjádření typu „zmenšení úbytku SASA“ mohou být matoucí, i zde uvedu rovnici popisující jednoznačně veličinu $\Delta\Delta SASA$

$$\Delta\Delta SASA = \Delta SASA_{WT} - \Delta SASA_{MUT}, \quad (2)$$

kde $\Delta\Delta SASA_{WT}$ je změna velikosti povrchu přístupného rozpouštědla nevázaných proteinů a jejich komplexu (bez vlivu mutací), $\Delta\Delta SASA_{MUT}$ je tatáž hodnota po mutaci rezidua na alanin a $\Delta\Delta SASA$ je potom rozdíl těchto dvou hodnot.

2.3 Množství dodnes testovaných reziduí

Výsledky jednotlivých experimentů ASM bohužel nejsou zapracovány do žádné z existujících proteinových databází, ale jsou shromažďovány odděleně v databázi Alanine Scanning Energetics Database (ASEdb). V dnešní době databáze obsahuje informace o 2915 mutovaných reziduích, zhruba dvě třetiny z nich se ale týkají výlučně různých forem lidského růstového hormonu. To v důsledku znamená, že ostatních 47 proteinů má průměrně uvedeno jen 23 mutovaných reziduí. Tato data jen dokreslují jak náročné ASM je a jaké naděje jsou vkládány do výpočetních metod predikce aktivních míst. [18]



Obrázek 5. Korelace $\Delta\Delta G$ a $\Delta\Delta SASA$. Převzato z [18]

Zajímavou skutečností je to, že kritéria pro $\Delta\Delta SASA$ a $\Delta\Delta G$ spolu jen špatně korelují [18], jak je vidět na obrázku 5. Tento nesoulad dosud nebyl uspokojivě vysvětlen. Z obou kritérií je však větší váha přikládána $\Delta\Delta G$.

3 Výpočetní metody predikce aktivních míst

Metody predikce aktivních míst v proteinech často vycházejí z metod pro predikci vazebných míst. Obě úlohy jsou si velmi podobné, ale zatímco vazebná místa jsou relativně velké oblasti na povrchu proteinů, aktivní místa jsou mnohem menší. Většina metod využívá pro predikci informaci o struktuře volných proteinů nebo jejich komplexů. Tyto metody se ukazují jako úspěšnější, mají však oproti metodám pracujícím čistě na základě informace ze sekvence jednu velkou nevýhodu. Množství proteinů s objasněnou strukturou je stále omezené. Odhaduje se, že prostorová struktura je známa jen asi u jednoho procenta známých interagujících proteinů [15]. Pro predikci aktivních míst je navíc často potřeba mít k dispozici modely s vysokým rozlišením, těch je samozřejmě ještě méně.

První a stále velice rozšířenou metodou určení struktury biomakromolekul, tedy i proteinů, je rentgenová spektroskopie označovaná též jako rentgenová krystalografie. Její využití však vyžaduje, aby zkoumaná látka tvořila krystal. Proteiny zpravidla krystalizují v podmínkách značně odlišných od těch fyziologických a nalezení správných podmínek pro krystalizaci je i dnes problematické [19]. Právě tato skutečnost představuje limitující faktor v získávání struktur dalších proteinů. Často se používají roztoky s velmi vysokou koncentrací solí. Vzhledem k tomu, že se nejedná o fyziologické podmínky, struktura proteinu v krystalu tak může být odlišná od jeho nativní struktury. Výstupem je difrakční obrazec, z něž se potom výpočetními technikami určuje rozložení hustoty elektronů ve vzorku a z něj následně jeho struktura.

Novější metoda, jejíž první aplikace proběhla již v padesátých letech minulého století [20], využívá nukleární magnetické rezonance (NMR). Není zde omezení pouze na molekuly tvořící krystaly, místo toho jsme limitováni velikostí molekuly [21]. Navíc je potřeba připravit vzorek v dostatečně vysoké koncentraci alespoň 0,2 mmol/l, který musí být stabilní po dostatečně dlouhou dobu (v řádu dnů) [22]. Omezení je sice menší, než u rentgenové krystalografie, výstupem zde ale není jednoznačná struktura. Výstupem je několik (řádově desítky) možných struktur odpovídajících naměřenému rozložení hustoty elektronů.

Tato práce je zaměřena na výpočetní metody predikce aktivních míst vycházející ze sekvence proteinů. Přesto začnu u metod vycházejících ze struktury proteinů a jejich komplexů a až potom detailněji rozeberu metody vycházející ze sekvence.

3.1 Metody založené na struktuře komplexu

3.1.1 MAPPIS

Vzhledem ke konzervativní povaze aktivních míst se jako logický postup jeví hledat je právě jako části proteinů zachované mezi členy různých proteinových rodin. Tuto myšlenku využívá algoritmus MAPPIS (*Multiple Alignment of Protein-Protein InterfaceS*) [13].

Pracuje na principu prostorového zarovnání proteinů, během něhož hledá optimální pozici podjednotek komplexu. Jde však ještě dále. Místo hledání konzervovaných reziduí hodnotí algoritmus povahu interakcí vznikajících během tvorby komplexu mezi rezidui vazebných míst obou reagujících proteinů. V jednotlivých komplexech potom hledá podobné vzorce těchto interakcí. Protože nejsou porovnávána jednotlivá rezidua, nýbrž jejich interakce, algoritmus je schopen detekovat aktivní místa i tam, kde proteiny nemají vzájemně zcela zachovalou sekvenci reziduí nebo terciární strukturu.

MAPPIS dosahuje senzitivity 66 % [13], jako referenční hodnotu bere v úvahu údaje prezentované v [23], kde je reziduum považováno za aktivní místo při $\Delta\Delta G \geq 1$ kcal/mol. Ke své činnosti algoritmus potřebuje dostatečné množství detailních prostorových struktur proteinů s podobnou funkcí, což může být omezující.

3.1.2 Hotpoint

Webserver Hotpoint využívá pro predikci aktivních míst několik parametrů. Jsou jimi $\Delta\Delta SASA$, rozdíl potenciálů mezi rezidui a vzájemná vzdálenost reziduí. Na základě těchto parametrů, je schopen s přednastaveným prahem dosáhnout senzitivity 59 % a pozitivní prediktivní hodnoty 73 % [24]. Klasifikační prahy však mohou být upraveny uživatelem v závislosti na jeho požadavcích, účinnost se tedy může při reálném použití lišit.

3.1.3 ROBETTA

V podstatě přímou aplikací experimentálních metod do prostředí výpočetní predikce je metoda používaná na serveru ROBETTA. Metoda je též někdy označována jako Robetta-Ala [25] a jde o výpočetní mutagenezi alaninovým skenováním. Autoři uvádějí, že metoda dosahuje senzitivity 79 % [26], jiné zdroje potom tvrdí, že při použití odlišných vstupních dat nejsou dosažené výsledky až tak dobré [9]. Přes snahu co nejvíce se přiblížit mutagenezi alaninovým vyhlazením, se hodnoty vypočítané $\Delta\Delta G$ od experimentální metody odlišují. Rozdíl je tak velký, že pro dosažení stejných výsledků, jakých ASM dosahuje s prahem 2 kcal/mol, zde musí být uvažována prahová hodnota 1 kcal/mol [26].

3.2 Metody založené na struktuře volných proteinů

3.2.1 pyDockNIP

Metod vycházejících ze struktury volných proteinů mnoho není. Popíši zde proto pouze tuto jednu. Metoda pyDockNIP je spojením metody pyDock [27] a metody popsané v [28]. PyDock byl navržen pro predikci vazebných míst proteinů pomocí jejich prostorového zarovnání (dokování). Dokování proteinů přitom zvládá rychleji než [28]. Metoda představená v [28] je již zaměřena na predikci aktivních míst a do syntézy s pyDockem přináší vlastní kritériální funkci NIP. Název pyDockNIP potom reprezentuje základní vlastnosti metody:

- py – metoda je realizována v jazyce Python,
- Dock – pracuje na principu dokování volných proteinů,
- NIP – jako kritérium pro klasifikaci aktivních míst je použita veličina *Normalized Interface Propensity* (NIP, normalizovaná četnost rezidua).

PyDockNIP hledá optimální prostorové zarovnání volných proteinů, přičemž jako kritérium používá hodnotu NIP. Ze souboru dat jsou náhodně vybírána rezidua tak, aby měl výsledný podsoubor stejnou distribuci SASA jako vazebné místo proteinu. Pro každé reziduum tohoto podsouboru je určena hodnota NIP. Ta je dána následující rovnicí

$$NIP = \log_2 \frac{w_i}{S_i}, \quad (3)$$

kde w_i je relativní četnost rezidua i v celém souboru dat a S_i je relativní četnost rezidua i ve vybraném podsouboru [28]. Za aktivní místa jsou považována rezidua s $NIP > 0,4$ [28]. Metoda potom dosahuje senzitivity 43 % a pozitivní prediktivní hodnoty 68 % [9]. Hodnota prahu však ani zde není pevně daná a při její volbě by měl uživatel brát v potaz, s jakým cílem analýzu vůbec provádí [28].

3.3 Metody založené na sekvenci proteinu

3.3.1 ISIS

Metoda ISIS (*Interaction Sites Identified from Sequence*) je další z metod původně navržených pro predikci vazebných míst v proteinech. Ojedinelá je v tom, že využívá pouze informaci ze sekvence proteinu, a to dokonce bez nutnosti znalosti vazebného partnera [15]. Nutno však podotknout, že během učení této neuronové sítě byly použity i informace o struktuře proteinů [9].

Když byla ISIS nasazena na řešení problému predikce vazebných míst, dokázala metoda identifikovat pouze malou část reziduí tvořících vazebná místa (viz obrázek 3). Byla to však rezidua, u nichž se předpokládala velká důležitost. Později se ukázalo, že

tyto výsledky dobře korelují s aktivními místy určenými pomocí ASM. Sami autoři potom přiznávají, že k predikci aktivních míst došlo omylem. Celkem je vyhodnocováno 189 parametrů. Přesto, že neuronová síť se naučila klasifikovat aktivní místa ze sekvence, dosažené výsledky bohužel nebylo možno interpretovat ve smyslu nalezení obecného kritéria k jejich klasifikaci mimo tuto síť. [15]

Při senzitivě 15 % dosahuje ISIS pozitivní prediktivní hodnoty 89 % [9]. Nastavení parametrů bylo přitom vybráno autory záměrně pro dosažení takto nízké senzitivity a vysoké pozitivní prediktivní hodnoty za účelem srovnání s jinými metodami [15]. Favorizovaným ukazatelem v takových srovnáních je totiž právě pozitivní prediktivní hodnota. Výsledky dosažené při jiném nastavení parametrů bohužel nebyly zveřejněny.

Důležité je zmínit i to, jak bylo takto vysoké pozitivní prediktivní hodnoty dosaženo. Za skutečná aktivní místa byla považována pouze ta rezidua, která při experimentálním ASM dosáhla $\Delta\Delta G \geq 2,5$ kcal/mol. Správná klasifikace reziduí neoznačených jako aktivní místa byla přiznána pouze těm, která při ASM dosáhla $\Delta\Delta G = 0$ kcal/mol [15]. Takto nastavené klasifikační prahy vůbec neuvažují rezidua s $\Delta\Delta G < 0$ a $0 < \Delta\Delta G < 2,5$ [9]. Následkem je problematické srovnání dosažených výsledků s jinými metodami uvažujícími celý rozsah volné vazebné energie.

3.3.2 Využití digitální filtrace

Metod pokoušejících se o nasazení postupů pro zpracování signálů na predikci aktivních míst proteinů se v poslední době objevilo několik. Zde popíši metodu publikovanou v článku [29].

Tabulka 1. Hodnoty EIIP pro jednotlivé aminokyseliny. Převzato z [30]

Aminokyselina	EIIP [Ry]	Aminokyselina	EIIP [Ry]
Alanin	0,0373	Kyselina glutamová	0,0058
Arginin	0,0959	Leucin	0,0000
Asparagin	0,0036	Lysin	0,0371
Cystein	0,0829	Methionin	0,0823
Fenylalanin	0,0946	Prolin	0,0198
Glutamin	0,0761	Serin	0,0829
Glycin	0,0050	Threonin	0,0941
Histidin	0,0242	Tryptofan	0,0548
Isoleucin	0,0000	Tyrozín	0,0516
Kyselina asparagová	0,1263	Valin	0,0057

Aby bylo možné použít metody digitální filtrace, je nejdříve nutné sekvenci proteinu skládající se z jednotlivých aminokyselin převést na číselné hodnoty.

Ramachandran zde pro číselnou reprezentaci aminokyselin zvolil parametr EIIP (*Electron-Ion Interaction Potencial*), dříve známý jako PEII (*Potencial of Electron-Ion Interaction*) [30]. Tato veličina reprezentuje průměrnou energii valenčních elektronů dané aminokyseliny [29]. Hodnoty pro jednotlivé aminokyseliny kódované v lidském genomu jsou uvedeny v tabulce 1. Veškerá literatura uvádí hodnoty EIIP v Rydberzích ($1 \text{ Ry} = 2,18 \cdot 10^{-18} \text{ J}$).

Vidíme, že hodnoty EIIP se pohybují v rozmezí 0 - 0,1263 Ry. Signál složený z těchto hodnot bude tedy mít výraznou hodnotu stejnosměrné složky. Ta by v dalším zpracování mohla přehlušit ostatní frekvenční složky, je proto žádoucí ji odstranit.

Tato metoda vyžaduje pro vstup hned několik příbuzných proteinů. Autoři uvádějí, že v závislosti na konkrétních rodinách mohou pro dobré výsledky stačit dva, nebo také devět [29]. V sekvencích těchto proteinů nahradí znaky aminokyselin hodnotami EIIP a určí se jejich konsenzuální spektrum. To se počítá podle následujícího vzorce [29]

$$S(e^{j\omega}) = |X_1(e^{j\omega}) \cdot X_2(e^{j\omega}) \cdot \dots \cdot X_M(e^{j\omega})|, \quad (4)$$

kde $X_i(e^{j\omega})$ je fourierovské spektrum proteinu i a $S(e^{j\omega})$ je konsenzuální spektrum všech zkoumaných proteinů.

V konsenzuálním spektru jsou viditelné výrazné peaky pro frekvence reprezentující společné funkce všech zkoumaných proteinů, většina ostatních frekvenčních složek má nulový, nebo velmi malý výkon. Frekvence s výrazným výkonem jsou označovány jako charakteristické frekvence. Příklad konsenzuálního spektra je možné vidět na obrázku 9 v kapitole 4.1.4.

Výpočet konsenzuálního spektra a jeho význam je jádrem modelu rezonančního rozpoznání (*Resonant Recognition Model*, RRM). Tento model předpokládá přítomnost stejné charakteristické frekvence jak u proteinu, tak u jeho vazebného partnera, ovšem pokaždé v opačné fázi, což je příčinou vzniku rezonance [25].

Dalším krokem je určit kde v sekvenci proteinů jsou tyto frekvence zastoupeny nejvíce. Tuto informaci není z fourierovského spektra možné přímo vyčíst [25]. Je proto navržen úzkopásmový filtr s cílem charakteristickou frekvenci ze signálu izolovat. Pro své vhodné vlastnosti, zvláště monotónní přenosovou charakteristiku v propustném pásmu, jsou používány inverzní Čebyševovy filtry. Jelikož tyto filtry jsou typu IIR (*Infinite Impulse Response*), výpočet jejich zpoždění by byl netriviální [29]. Využívá se proto techniky známé jako *zero-phase filtering*, kdy je signál prohnán filtrem dvakrát, přičemž podruhé signál do filtru vstupuje pozpátku.

Ze získaného signálu se vypočítá jeho energie, přičemž získaná posloupnost hodnot již koresponduje s jednotlivými rezidui. Autoři tuto posloupnost nazývají energetickou sekvencí. Aktivní místa jsou detekována tam, kde peak energetické

sekvence přesáhne hodnotu práhu. Práh je určen jako násobek průměrné energie všech reziduí sekvence. Přičemž Ramachandran ve své práci jako práh používá přímo tuto průměrnou energii. Za aktivní místo nepovažuje všechny hodnoty přesahující práh, ale pouze vrcholy dostatečně vysokých peaků.

Autoři bohužel nezveřejnili žádné hodnoty týkající se úspěšnosti metody. Z textu jejich zprávy je patrné, že se soustředili především na minimalizaci výpočetní náročnosti, která je podle nich zhruba pětinasobně nižší než u jejich dříve zveřejněné metody publikované ve článku [31].

3.3.3 Využití S-transformace

Metoda publikovaná v [25], vychází stejně jako předchozí popisovaná metoda z RRM. Snaží se však adresovat nedostatky Fourierovy transformace při popisu nestacionárního signálu, jakým sekvence proteinu je. Řešení tohoto problému nabízejí transformace používané pro časově frekvenční analýzu signálu. Například krátkodobá Fourierova transformace (*short time Fourier transform*, STFT), nebo spojitá vlnková transformace (*continuous wavelet transform*, CWT). Oba tyto postupy ale trpí nedostatky, které by měla překonat S-transformace (ST) poprvé popsána ve článku [32].

S-transformace signálu $x(t)$ je popsána rovnicí

$$S(\tau, f) = \int_{-\infty}^{\infty} x(t) \omega(\tau - t, f) e^{-j2\pi ft} dt, \quad (5)$$

kde $x(t)$ je časově proměnný signál na vstupu transformace, $S(\tau, f)$ je vzorek ST pro frekvenci f a posunutí τ okna ω o [25]. Okno ω je přitom dáno jako

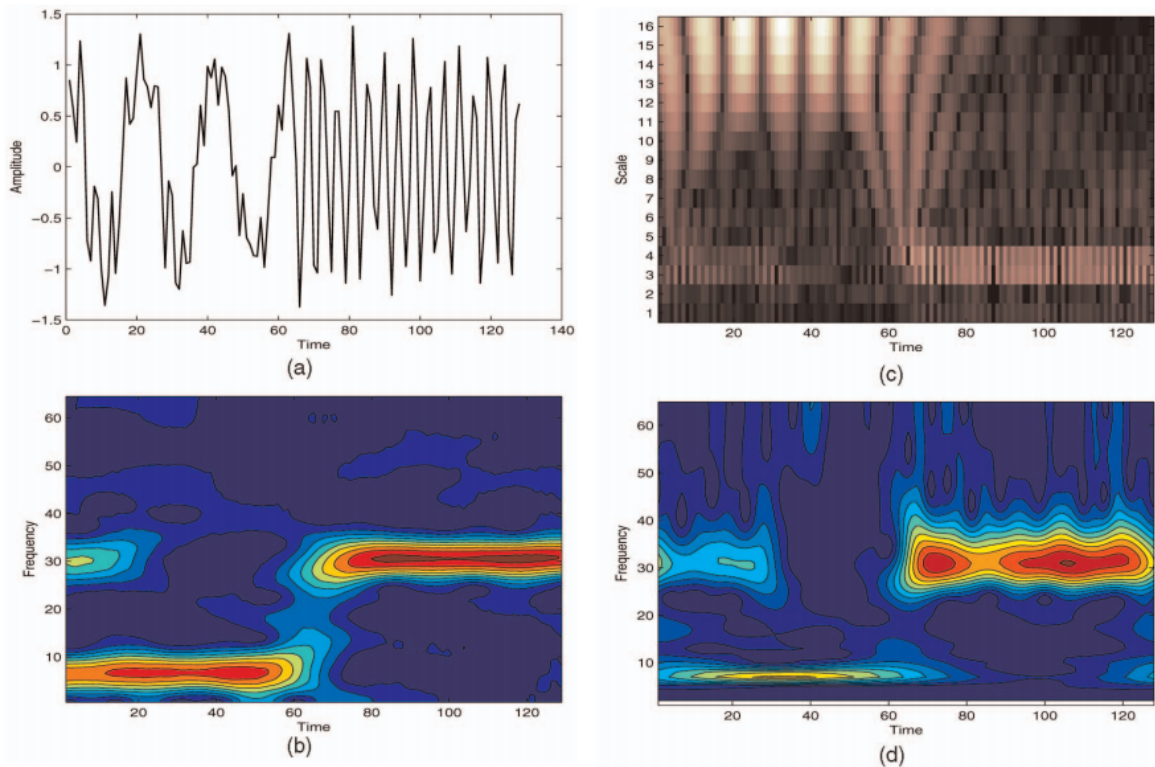
$$\omega(t, \sigma) = \frac{1}{\sigma(f)\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2(f)}}, \quad (6)$$

přičemž šířka σ okna ω je

$$\sigma(f) = \frac{1}{|f|}. \quad (7)$$

Právě skutečnost, že šířka okna je závislá na frekvenci poskytuje ST výhodu oproti STFT, kde je časové rozlišení na všech frekvencích konstantní. Sahu demonstruje srovnání STFT, CWT a ST na signálu, v jehož první polovině se nachází sinusovka o frekvenci 6 Hz a ve druhé polovině 30 Hz. Celý signál je potom ještě zašuměn rušením o intenzitě 0 dB. Tento signál, spolu se spektry vypočítanými pomocí všech tří metod můžeme vidět na obrázku 6. Zatímco spektrum poskytnuté metodou CWT je od zbylých dvou jasně odlišné, spektra STFT a ST vypadají na první pohled velice podobně. Při bližším pohledu si můžeme všimnout, že spektrum ST daleko přesněji vymezuje nižší frekvenci v signálu (6 Hz), přičemž rozlišení na frekvenci 30 Hz už je

u obou metod srovnatelné. Porovnání s CWT je náročnější, spektrum ST však zcela jasně vyniká ve své jednodušší interpretaci. Kromě toho všeho je ST invertibilní, to umožňuje nasazení časově proměnné filtrace, což by nám použití STFT nedovolilo. [25]



Obrázek 6. Srovnání STFT, CWT a ST; (a) Signál, (b) STFT spektrum, (c) CWT spektrum, (d) ST spektrum. Převzato z [25]

Tato metoda využívá stejného postupu jako metoda využívající digitální filtrace, a to až do kroku výpočtu konsenzuálního spektra včetně. Kromě konsenzuálního spektra je ale určeno i spektrum ST. Na rozdíl od konsenzuálního spektra, do výpočtu ST vstupuje pouze jedna sekvence, a to protein, který nás zajímá.

Takto získané ST spektrum se pronásobí s konsenzuálním spektrem. Potlačíme tak výkon na nežádoucích frekvencích, tedy v podstatě šum [25]. Následně je navržen časově proměnný filtr k izolaci charakteristických frekvencí. Ten je aplikován na sekvenci zajímavého proteinu. Z výstupního signálu filtru se spočítá energie stejně jako v předchozí metodě a stejně tak se vyhodnocuje. Využitím ST dosáhli autoři senzitivity 83,33 % a pozitivní prediktivní hodnoty 62,5 % [25].

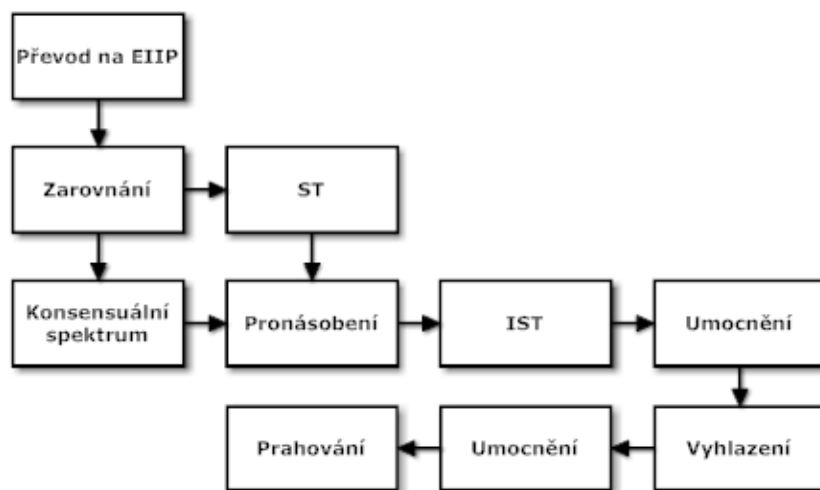
4 Navrhovaný algoritmus

Navrhovaný algoritmus predikce aktivních míst v proteinech vychází z metody popisované v kapitole 3.3.3 a článku [25]. Stejně jako tam, i zde je využita S-transformace a veličina EIIP k převodu sekvence reziduí na numerický signál. Opět jsou tedy využity hodnoty z tabulky 1. Během své práce jsem vyvinul několik podob algoritmu, ze kterých není vždy možné vybrat obecně tu nejlepší. V této kapitole proto nejdříve popíšu obecnou podobu algoritmu, označovanou později také jako metoda A, a poté jeho jednotlivé mutace.

4.1 Obecný popis algoritmu, metoda A

Algoritmus ve své podstatě vyhledává motivy, které v sekvencích proteinů během evoluce zůstaly zachované. Nevystačí si zde proto pouze se sekvencí zkoumaného proteinu, ale potřebuje ještě několik sekvencí proteinů příbuzných, aby bylo odkud čerpat informace o evoluci. Ideálním řešením se ukázalo srovnávat zkoumaný protein s proteiny stejné funkce, exprimovanými jinými organizmy.

Obrázek 7 ukazuje blokové schéma obecného algoritmu. Ten představuje základ, od kterého se v dalších kapitolách budou odvíjet jednotlivé jeho mutace. V dalším popisu potom tento základní algoritmus bude označován jako metoda A.



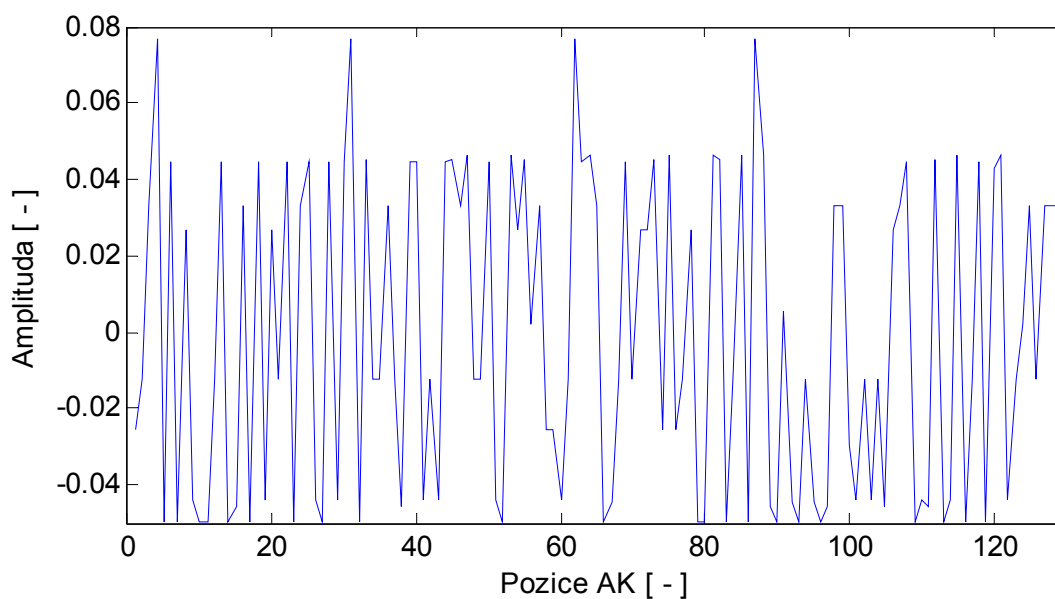
4.1.1 Vstupní data

Pokud hledáme sekvence proteinů v různých volně přístupných databázích, můžeme často najít záznamy obsahující celé geny, které v jedné sekvenci prezentují současně sekvenci proteinu i jeho signálního peptidu. Příkladem takové databáze je Uniprot [33]. Pro účely predikce aktivních míst je však nutné pracovat pouze se sekvencí samotného proteinu v podobě, v jaké se účastní interakcí v organismu. Tuto

sekvenci lze získat i z databáze Uniprot, zde ale byla použita data poskytnutá databází PDB [1]. Sekvence proteinů, se kterými je zkoumaný protein porovnáván, už potom pochází z Uniprotu. Proteiny se mezi jednotlivými organismy mohou ve své délce mírně lišit, je proto dobré mít v tomto srovnání i rezidua z jejich signálního peptidu, případně jiných produktů, se kterými v sekvenci sousedí. Důsledkem těchto dvou odlišných zdrojů jsou potom dva formáty identifikátorů. Sekvence zkoumaných proteinů (zpravidla lidských) zde budou identifikovány ve formátu PDB, zatímco sekvence jim příbuzných proteinů ve formátu databáze Uniprot.

4.1.2 Převod na hodnoty EIIP

Prvním krokem předzpracování sekvencí je jejich konverze na numerický signál. K tomu je využito veličiny EIIP. Jelikož tato veličina popisuje průměrnou energii valenčních elektronů každého rezidua, skládá se výsledný signál pouze z nezáporných hodnot. Tato skutečnost by se v jeho fourierovském spektru projevila výrazným peakem na nulové frekvenci, což je nežádoucí. Aby byl tento artefakt odstraněn, je od každého signálu odečtena průměrná hodnota EIIP napříč všemi aminokyselinami. Výsledný signál vidíme na obrázku 8.



Obrázek 8. Signál lidského interleukinu 4 po předzpracování

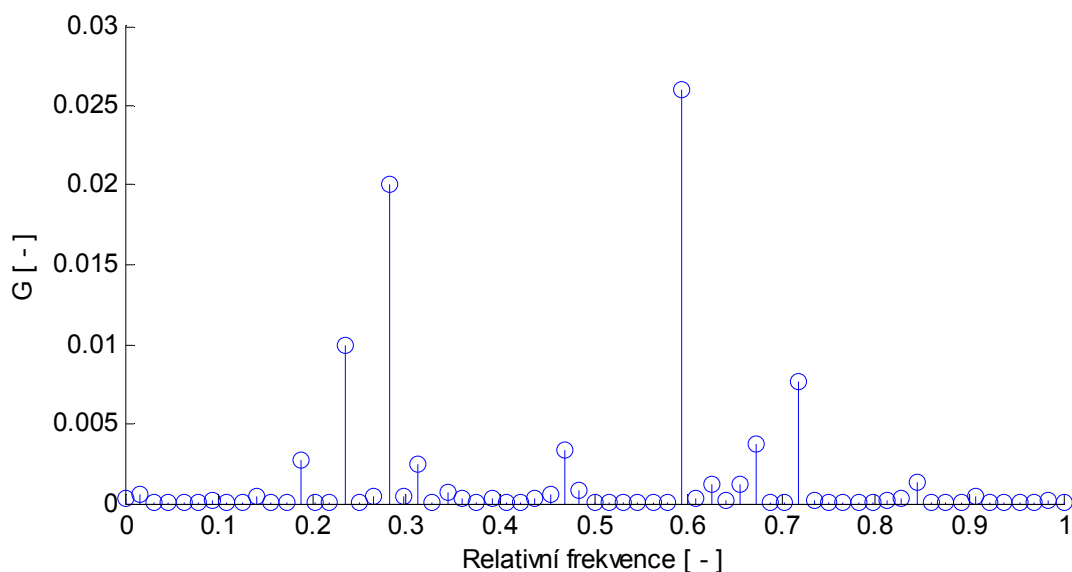
4.1.3 Zarovnání

Dále je třeba sekvence zarovnat. Není zde však řeč o zarovnání v klasickém bioinformatickém smyslu. Cílem je vybrat takové úseky sekvencí příbuzných proteinů, které se nejlépe shodují se sekvencí zkoumaného proteinu. Přitom je nežádoucí, aby do tohoto zarovnání byly jakkoli zařazeny mezery v kterékoli ze sekvencí. To by ovlivnilo frekvenční složení jejich signálu a zkreslilo celou analýzu. Používám proto jednoduché kritérium minimální Hammingovy vzdálenosti.

Díky způsobu získávání sekvencí, popsaném výše, by sekvence zkoumaného proteinu měla být vždy kratší (případně stejně dlouhá), než sekvence jemu příbuzných proteinů. Jakékoli ořezávání se proto ideálně uskutečňuje pouze u příbuzných sekvencí.

4.1.4 Výpočet konsensuálního spektra

Ze zarovnaných signálů je potom násobením jejich Fourierových spekter počítáno konsensuální spektrum podle rovnice 4. Příklad takového spektra pro jeden ze zkoumaných proteinů – konkrétně interleukin 4 – je uveden na obrázku 9.



Obrázek 9. Konsensuální spektrum interleukinu 4

Vidíme, že při výpočtu konsensuálního spektra došlo k vyrušení většiny frekvenčních složek. Jen několik zbylých složek má stále dostatečný výkon, aby ve spektru tvořily výrazné peaky. Tyto frekvence reprezentují společné funkce použitých proteinů a tedy i informace o vysoce konzervativních reziduích. Pro algoritmus je důležitá frekvence s největším výkonem (a tím pádem i funkce mezi druhy nejvíce zachovaná). Ta bude dále označována jako charakteristická frekvence. Fourierovské spektrum však neumožňuje určit, kde v signálu se tato frekvence nachází, k tomu je potřeba nasadit jednu z metod časově frekvenční analýzy.

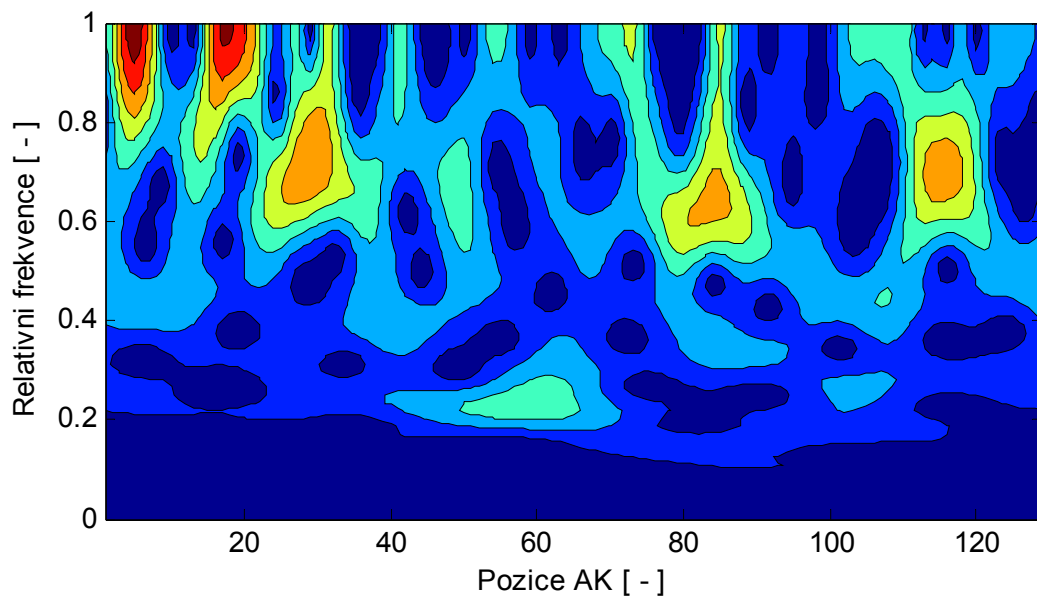
4.1.5 S-transformace

Časově frekvenční analýzu zde zprostředkovává právě S-transformace. Její výstup je velmi podobný STFT, ST však drží jednu nespornou výhodu. Je invertibilní. Na rozdíl od jejího použití v metodě podle [24] však používám odlišnou definici, ta se ukázala jednodušší pro implementaci. Uvádím diskrétní vzorec pro dopřednou (8) i zpětnou (9) transformaci:

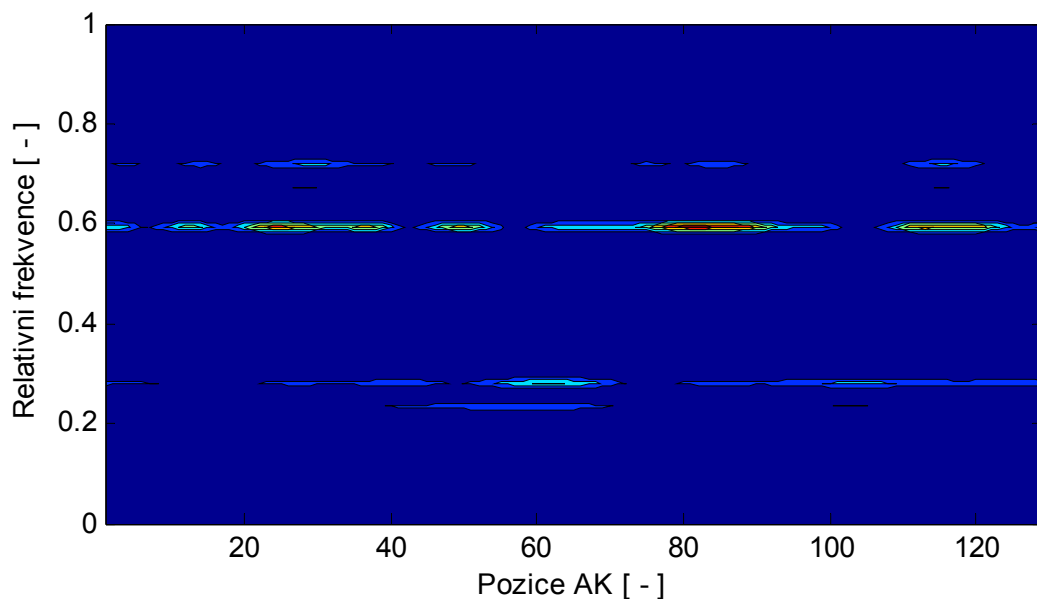
$$S(p, m) = \sum_{q=-M/2}^{M/2-1} U\left(\frac{m+q}{M}\right) \cdot e^{-2\left(\frac{\pi q}{M}\right)^2} \cdot e^{2i\pi \frac{qp}{M}} \quad (8)$$

$$u(n) = \frac{1}{M} \sum_{m=-M/2}^{M/2-1} \sum_{p=0}^{M-1} S(p, m) \cdot e^{\frac{2imn\pi}{M}} \quad (9)$$

$S(p, m)$ zde představuje koeficient S-transformace v čase p pro frekvenci m ; U je spektrum Fourierovy transformace diskrétního signálu $u(n)$ a M je počet vzorků signálu. Za proměnné p a m je v tomto případě dosazován index frekvence (případně času), nikoli jejich skutečná hodnota. [34]



Obrázek 10. ST spektrum lidského interleukinu 4



Obrázek 11. ST spektrum lidského interleukinu 4 po pronásobení s konsenzuálním spektrem

Obrázek 10 představuje typické ST spektrum proteinu. Na první pohled se může zdát, že žádné relevantní údaje neobsahuje. Je však vhodné povšimnout si, že většina výkonu se odehrává na vysokých frekvencích. To si lze ověřit i při pohledu na samotný signál na obrázku 8. Přitom se jedná o vlastnost, kterou mají všechny proteiny společnou.

4.1.6 Násobení ST spektra a konsenzuálního spektra

Pro další práci je ST spektrum pronásobeno s konsenzuálním spektrem. Prakticky se tak odfiltruje výkon na většině frekvencí. Zůstanou jen pruhy odpovídající peakům v konsenzuálním spektru. Výsledek této operace můžeme pozorovat na obrázku 11.

Je vidět, že ze zbylých pruhů ten nejvýraznější přesně odpovídá charakteristické frekvenci z konsenzuálního spektra, tedy 0,6. To je frekvence, se kterou bude algoritmus pracovat i v dalším kroku.

4.1.7 Inverzní S-transformace a zpracování jejího výstupu

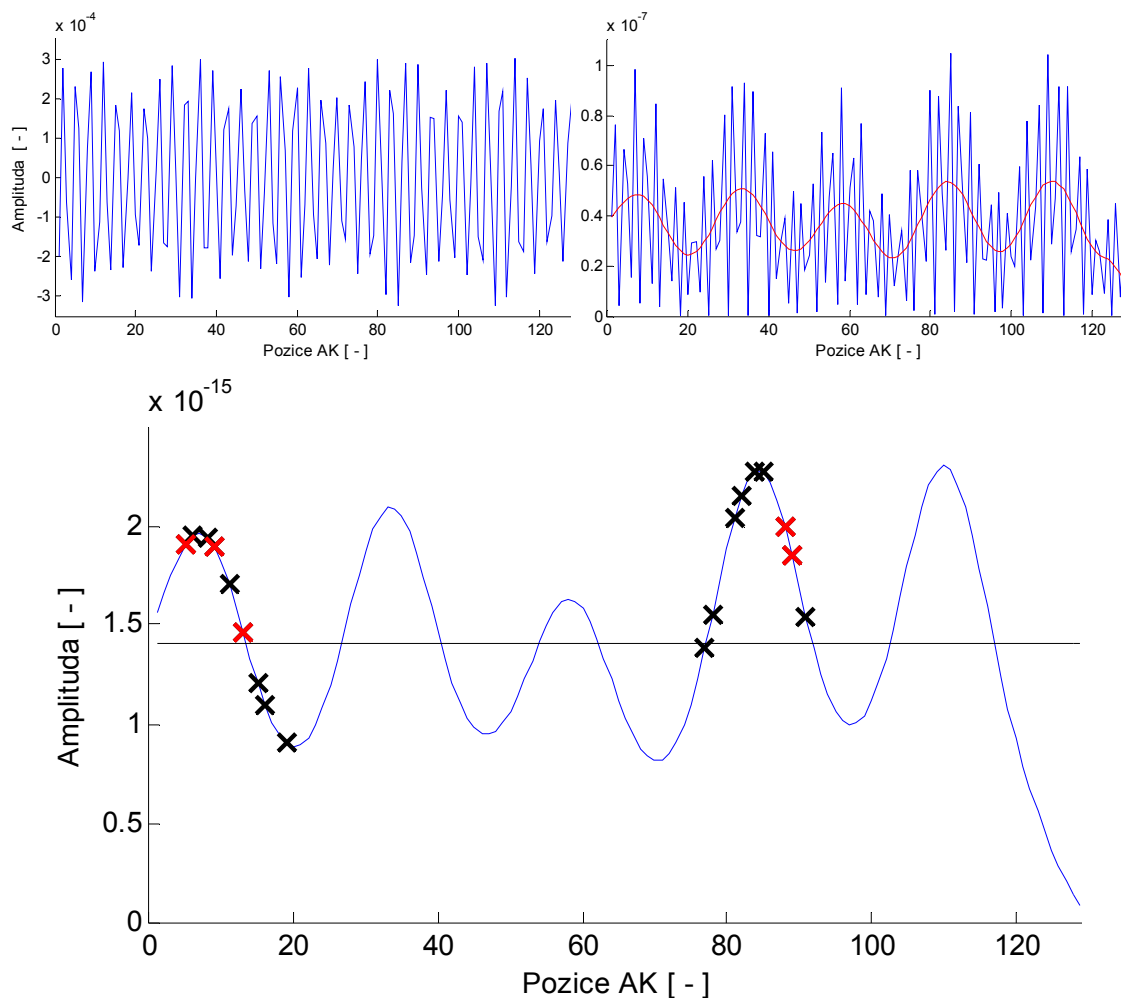
Před aplikací inverzní S-transformace je spektrum dále filtrováno. Filtrace je provedena jednoduchým násobením spektra s maskou složenou z nul a jedniček, jde tedy o 2D obdobu nulování spektrálních čar. Střed propustného pásma leží na charakteristické frekvenci a jeho šířka odpovídá pětině frekvenčního rozsahu.

Signál na výstupu IST na první pohled nenese mnoho užitečných informací, proto je potřeba ho dále upravit. Signál je umocněn na druhou a vyhlazen. Pro vyhlazení používáme jednoduchou dolní propust o řádu 20 s mezní frekvencí 0,1. Takto vyhlazený signál je opět umocněn na druhou a tím konečně dostáváme signál, na němž bude pomocí prahování prováděna detekce aktivních míst. Výsledky zmiňovaných úprav jsou vykresleny na obrázku 12.

4.1.8 Prahování

Práh znázorněný na obrázku 12 je v tomto případě roven průměrné hodnotě výsledného signálu. V dalších kapitolách je však jeho hodnota posouvána za účelem dosažení optimálních výsledků. Posunutí je vždy vyjádřeno násobkem směrodatné odchylky signálu, o niž je práh od jeho průměrné hodnoty odchýlen. Výše prahu je potom vyjádřena právě tímto násobkem.

Všechna rezidua, pro něž se křivka nachází nad prahem, algoritmus detekuje jako aktivní místa. Vidíme, že v tomto konkrétním případě se nám podařilo úspěšně detekovat všech pět známých aktivních míst. Kromě nich došlo bohužel i k detekci devíti z třinácti reziduí, o nichž je známo, že aktivními místy nejsou. Konkrétní ukazatele úspěšnosti jsou vyčísleny v kapitole 6.



Obrázek 12. Vlevo nahoře - výstup IST; vpravo nahoře - výstup IST po umocnění na druhou (modře) a po vyhlazení (červeně); dole - vyhlazený signál po opětovném umocnění (modře), aktivní místa dle ASEdb (červené křížky), residua v ASEdb neklasifikovaná jako aktivní místa (černé křížky), detekční práh našeho algoritmu (černá linie)

4.2 Metoda B

Postup metody B je znázorněn na obrázku 13. Oproti metodě A je o poznání jednodušší a přímější. ST i konsenzuální spektrum se zde počítá úplně stejně, už ale není třeba tyto dva prvky spolu násobit.

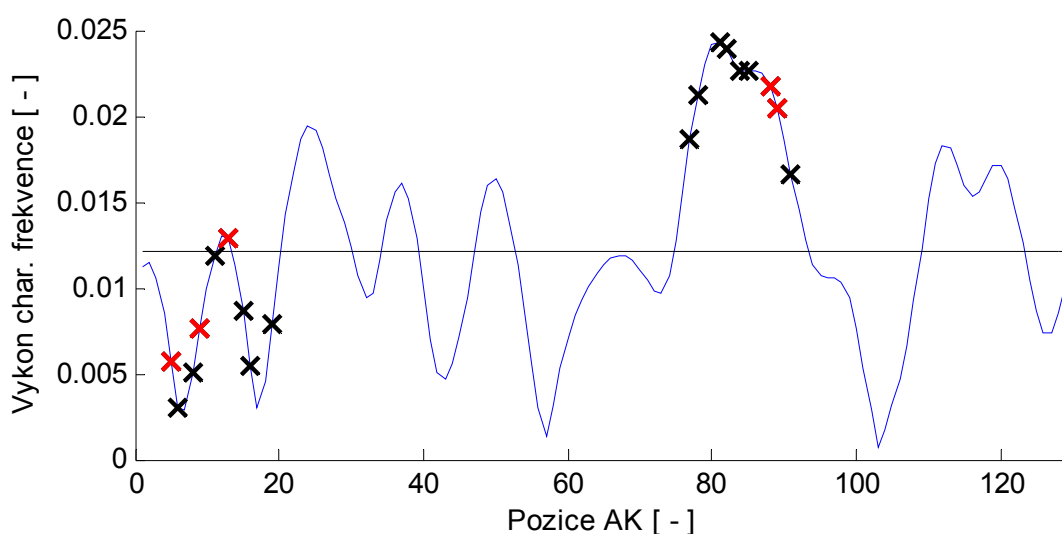


Obrázek 13. Metoda B – blokové schéma

4.2.1 Charakteristická frekvence

Konsenzuální spektrum zde slouží jedinému účelu, a to k nalezení charakteristické frekvence. Tato frekvence opět reprezentuje společnou funkci všech proteinů, která se mezi jednotlivými organismy zachovala nejlépe. I při popisu této metody se budu držet příkladu interleukinu 4, charakteristická frekvence je proto opět 0,6, jak bylo ukázáno už na obrázku 9.

Pro další práci bude použit řez ST spektra právě na charakteristické frekvenci. Tento signál nám udává průběh výkonu dané frekvence v závislosti na pozici v sekvenci. Dá se proto očekávat, že pozice, kde má charakteristická frekvence vysoký výkon, budou právě hledanými aktivními místy.

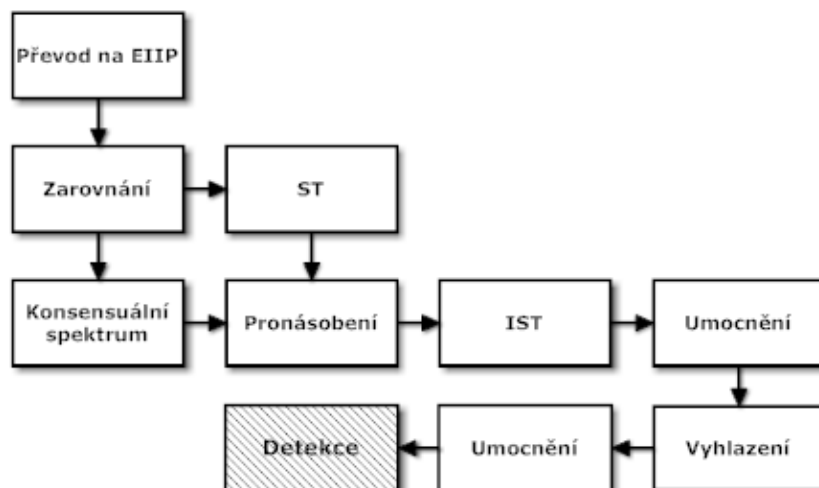


Obrázek 14. Metoda B - průběh charakteristické frekvence interleukinu 4 (modře), práh (černě), aktivní místa podle ASEdb (červené křížky), rezidua podle ASEdb neklasifikovaná jako aktivní místa (černé křížky)

Na obrázku 14 vidíme výsledek detekce pomocí metody B. Jako aktivní místa jsou opět detekována nadprahové vzorky. Práh je v tomto příkladě roven průměrné hodnotě signálu. S takovým nastavením se povedlo úspěšně detekovat tři z pěti známých aktivních míst. Falešně bylo detekováno i sedm z třinácti reziduí, o nichž je známo, že aktivními místy nejsou.

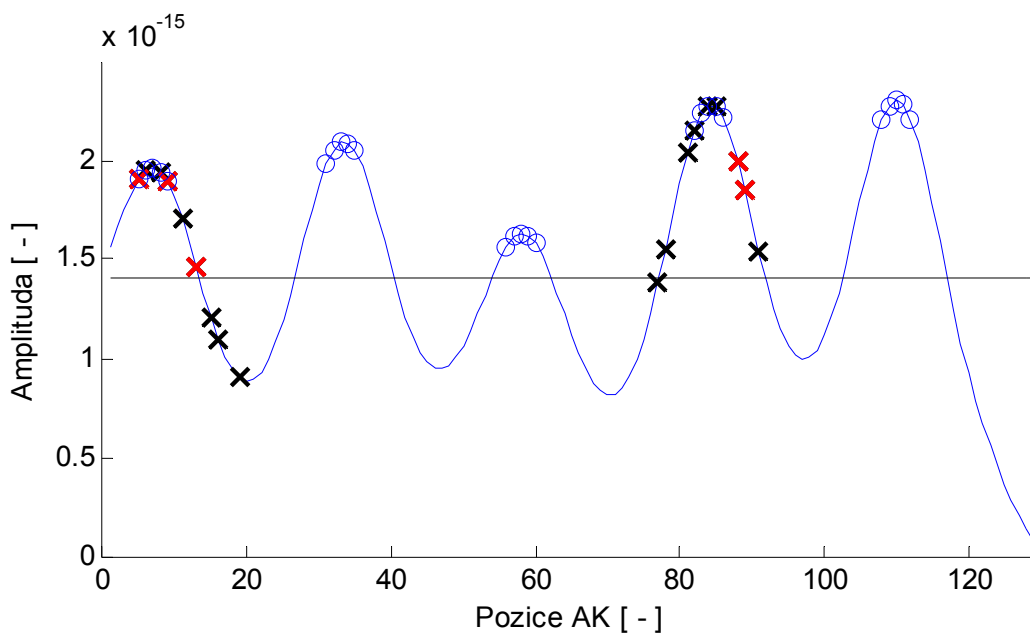
4.3 Metoda C

Tato metoda využívá pro detekci signál zpracovaný metodou A, odlišný je však způsob detekce. Detekují se zde pouze vrcholy všech nadprahových peaků a určitý počet vzorků v jejich okolí. V důsledku to znamená, že vzorek detekovaný jako aktivní místo může ve skutečnosti ležet pod stanoveným prahem, to ovšem ničemu nevádí.



Obrázek 15. Metoda C – blokové schéma

Jelikož je tento postup velice podobný postupu metody A, jediným rozdílem ve schématu je nahrazení bloku *Prahování* blokem *Detekce*. Tyto pojmy by neměly být zaměňovány. Pro zvýraznění změny je tento blok na obrázku 15 odlišen šrafovou.

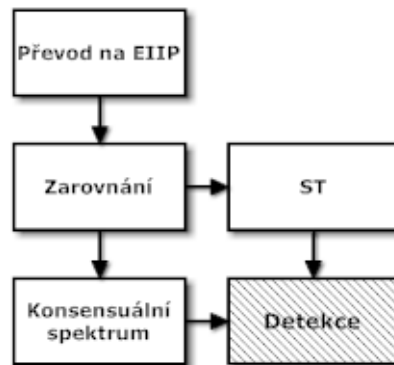


Obrázek 16. Metoda C – detekce; vyhlazený a umocněný výstup IST (modře), práh (černě), aktivní místa dle ASEdb (červené křížky), rezidua dle ASEdb neklasifikovaná jako aktivní místa (černé křížky), rezidua algoritmem detekovaná jako aktivní místa (kroužky)

Obrázek 16 ukazuje výsledek dosažený touto metodou. Na rozdíl od předchozích metod jsou zde detekovaná aktivní místa vyznačena kroužkem. Jedná se vždy o dva vzorky od vrcholu detekovaného peaku na každou stranu. Touto metodou se podařilo úspěšně detekovat dvě z pěti známých aktivních míst a falešně detekovat pět z třinácti reziduí, o nichž je známo, že aktivními místy nejsou.

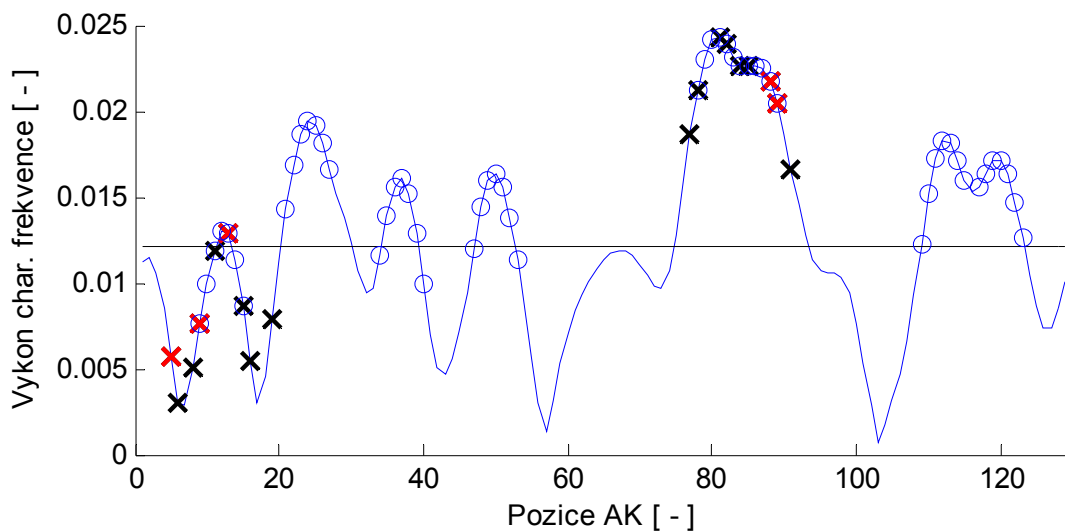
4.4 Metoda D

Tato metoda detekuje aktivní místa v řezu ST spektra stejně jako metoda B. Detekce ale probíhá způsobem představeným v metodě C. I zde se detekují vrcholy peaků a jako aktivní místa jsou považovány vzorky poblíž těchto peaků. Blokové schéma je znázorněno na obrázku 17.



Obrázek 17. Metoda D – blokové schéma

Příklad výstupu této metody je znázorněn na obrázku 18. Práh je zde roven opět průměrné hodnotě signálu a jako aktivní místa jsou detekována rezidua vzdálená od vrcholu nejvýše tři vzorky. Tyto body jsou v grafu označeny kroužkem. Na interleukinu 4 se podařilo úspěšně detekovat čtyři z pěti známých aktivních míst a sedm ze třinácti reziduí, která podle ASEdb aktivními místy nejsou.



Obrázek 18. Metoda D – detekce; řez ST spektrem na charakteristické frekvenci (modře), práh (černě), aktivní místa podle ASEdb (červené křížky), rezidua podle ASEdb neklasifikovaná jako aktivní místa (černé křížky), rezidua detekovaná algoritmem jako aktivní místa (kroužky)

Kromě toho je zde vidět, že algoritmus jako aktivní místa detekuje i ty vzorky, které se nacházejí pod prahem, jak již bylo zmíněno u předchozí metody. Ovšem jen

za předpokladu, že jsou tyto vzorky dostatečně blízko vrcholu nadprahového peaku. Nejméně v jednom případě je zde tato detekce oprávněná.

5 Pseudokód a vývojové diagramy

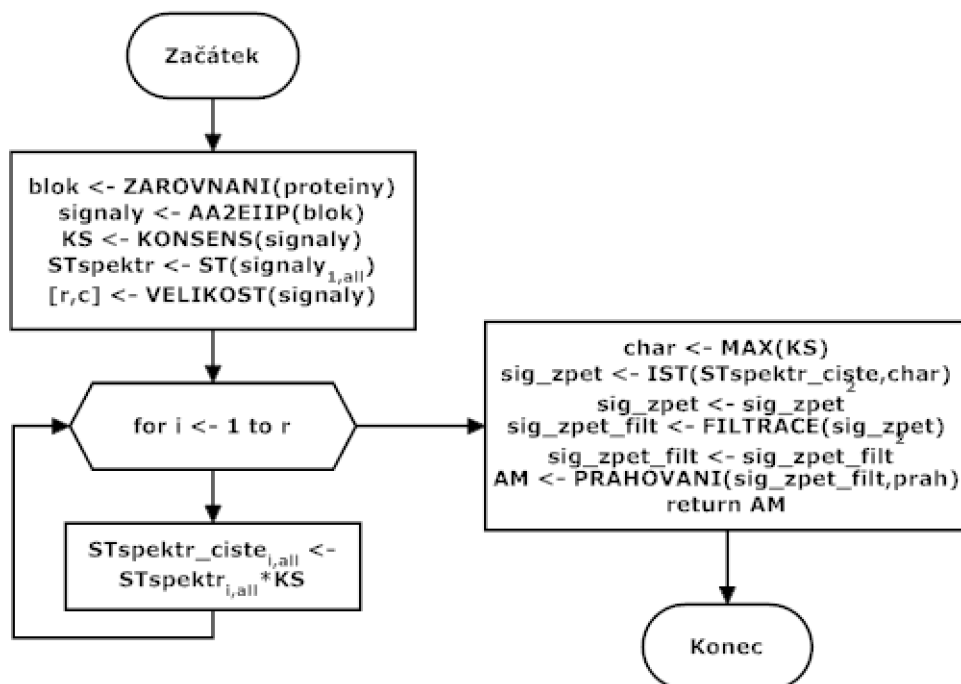
V této kapitole bude uveden pseudokód k některým z funkcí metody A popsané v minulé kapitole. Ke stejným funkcím zde budou uvedeny i vývojové diagramy. Funkce PREDIKCE_AM reprezentuje prakticky celou metodu A; AK_NA_EIIP zajišťuje převod sekvence na hodnoty EIIP a HAMMING hledá vzájemné posunutí dvou sekvencí, při kterém mají nejmenší vzájemnou Hammingovu vzdálenost.

PREDIKCE_AM(proteiny,prah)

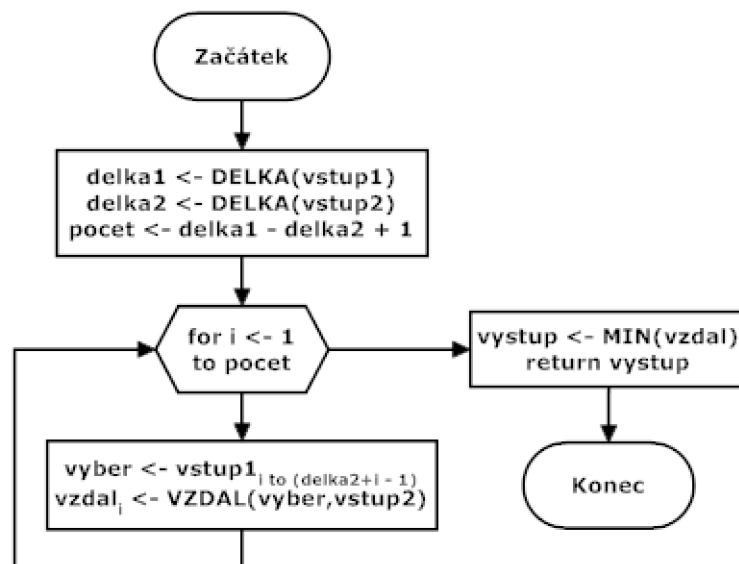
```
1  blok ← ZAROVNANI(proteiny)
2  signaly ← AA2EIIP(blok)
3  KS ← KONSENS(blok)
4  STspektr ← ST(signaly1,all)
5  [r,c] ← VELIKOST(signaly)
6  for i ← 1 to r
7  STspektr_cistei,all ← STspektri,all*KS
8  char ← MAX(KS)
9  sig_zpet ← IST(ST,STspektr_ciste,char)
10 sig_zpet ← sig_zpet2
11 sig_zpet_filt ← FILTRACE(sig_zpet)
12 sig_zpet_filt ← sig_zpet_filt2
13 AM ← PRAHOVANI(sig_zpet_filt,prah)
14 return AM
```

HAMMING(vstup1,vstup2)

```
1  delka1 ← DELKA(vstup1)
2  delka2 ← DELKA(vstup2)
3  pocet ← delka1 – delka2 + 1
4  for i ← 1 to pocet
5      vyber ← vstup1i to (delka2+i-1)
6      vzdal(i) ← VZDAL(vyber,vstup2)
7  vystup ← MIN(vzdal)
8  return vystup
```



Obrázek 19. Vývojový diagram pro funkci NAJDI_AM

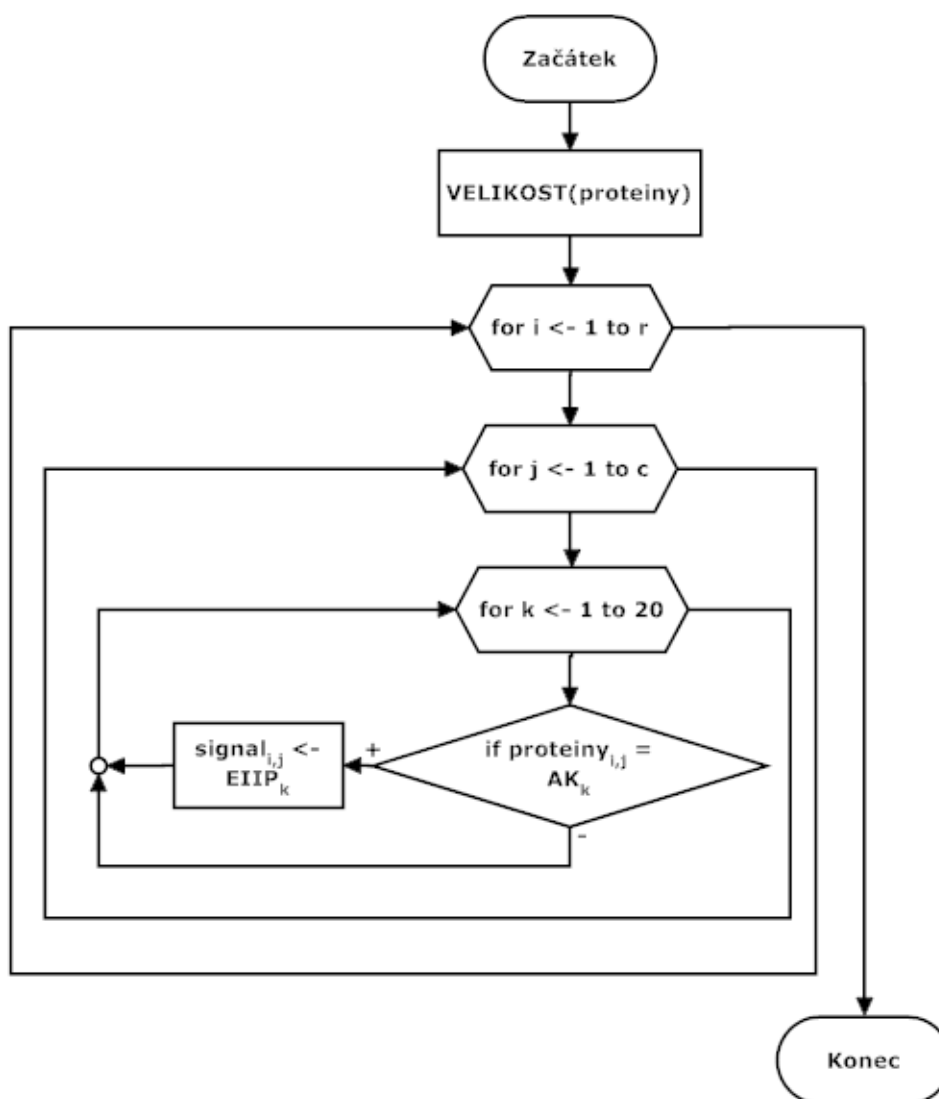


Obrázek 20. Vývojový diagram pro funkci HAMMING

```

AK_NA_EIIP(prot, EIIP, AK)
1  [r, c] ← VELIKOST(proteiny)
2  for i ← 1 to r
3      for j ← 1 to c
4          for k ← 1 to 20
5              if proteinyi,j = AKk
6                  signali,j ← EIIPk
7  return signal

```



Obrázek 21. Vývojový diagram pro funkci AA_NA_EIIP

V pseudokódu používám funkce, jejichž účel nemusí být na první pohled zřejmý, proto je zde pro jistotu popíšu:

Funkce ZAROVNANI zajišťuje zarovnání sekvencí tak, že z delší sekvence je vybrána pouze její část, délkou odpovídající kratší sekvenci, a to tak, aby výsledné dvě sekvence měly co nejmenší Hammingovu vzdálenost. Podfunkcí této funkce je funkce HAMMING.

Funkce KONSENS vypočítá konsenzuální spektrum ze signálů, které jsou jí dány na vstup, podle vzorce 4. Vstupem je matice, která na každém svém řádku nese jeden signál.

Funkce ST počítá S-transformaci signálu na vstupu podle vzorce 5. Na rozdíl od DFT je zde vstupem jediný signál, a to signál reprezentující analyzovaný protein. Tento signál je v matici signálů vždy uveden na prvním řádku.

Funkce VELIKOST určí velikost vstupní proměnné, první proměnná na výstupu potom určuje počet řádků matice, druhá proměnná počet sloupců.

Funkce MAX uloží do svého výstupu pozici maximálního prvku ve vektoru na jejím vstupu.

Funkce IST provádí inverzní S-transformaci a to včetně filtrace nulováním spektrálních čar v okolí charakteristické frekvence.

Funkce FILTRACE filtruje vstup pomocí dolní propusti o řádu 20 a mezní frekvenci 0,1.

Funkce PRAHOVANI provádí prahování. Prvním vstupem je signál, na němž bude prahování prováděno. Druhým vstupem je hodnota prahu uvedená jako násobek směrodatné odchylky signálu z prvního vstupu od jeho průměru.

Funkce DELKA uloží do výstupu délku vektoru na vstupu.

Funkce VZDAL počítá Hammingovu vzdálenost svých dvou vstupů.

Funkce MIN do své výstupní proměnné uloží pozici nejmenšího prvku ze vstupního vektoru.

Pro úplnost uvádím i popis významu důležitých proměnných:

- **AK** - znaky reprezentující jednotlivé aminokyseliny, musí být uvedeny ve stejném pořadí jako jejich EIIP v proměnné EIIP,
- **AM** - pozice aktivních míst, výstup funkce NAJDI_AM,
- **blok** – zarovnané sekvence proteinů,
- **delka1** – délka vstupu jedna funkce HAMMING,
- **delka2** – délka vstupu jedna funkce HAMMING,
- **EIIP** - EIIP jednotlivých aminokyselin, musí být uvedeny ve stejném pořadí jako jejich znaky v proměnné AK,
- **char** – index charakteristické frekvence,
- **KS** – konsenzuální spektrum,
- **pocet** – nejvyšší možný počet vzájemných posunutí vstupů funkce HAMMING tak, aby se stále překrýval stejný počet vzorků,

- **prah** – hodnota prahu uvedená jako násobek směrodatné odchylky prahovaného signálu přičtený k jeho průměru,
- **proteiny** – vstupní sekvence proteinů,
- **sig_zpet** – signál získaný inverzní S-transformací, případně jeho druhá mocnina,
- **sig_zpet_filt** – *sig_zpet* po filtraci funkcí FILTRACE, případně jeho druhá mocnina,
- **signal** - sekvence proteinu převedená na numerické hodnoty,
- **signaly** – matice sekvencí všech proteinů převedených na numerické hodnoty,
- **STspektr** – spektrum S-transformace,
- **STspektr_ciste** – spektrum S-transformace po násobení s konsenzuálním spektrem,
- **vstup1** – první vstup funkce HAMMING,
- **vstup2** – druhý vstup funkce HAMMING, v tomto vstupu se nachází signál zkoumaného proteinu,
- **vyber** – část prvního vstupu funkce HAMMING o stejné délce jako druhý vstup,
- **vystup** – index prvního vzorku proměnné *vstup1*, kterým začíná úsek s nejmenší Hammingovou vzdáleností s druhým vstupem.

6 Dosažené výsledky

6.1 Referenční data

Pro ověření účinnosti algoritmu bylo vybráno osm proteinů, které byly již dříve podrobeny ASM, a výsledky z těchto experimentů jsou uloženy v ASEdb. Jak už bylo popsáno v kapitole 2, ASM je finančně i časově velice náročná laboratorní technika. Z toho důvodu ani u vybraných proteinů nejsou parametry $\Delta\Delta G$ a $\Delta\Delta SASA$ známy u všech jejich reziduí. Obecné informace o těchto proteinech je možno nalézt v tabulce 2.

Tabulka 2. Základní informace o referenčních proteinech

Název proteinu	PDB kód	Délka proteinu [rezidua]	ASEdb (aktivní místa/celkem)
Barnase	1brs	110	6/8
Coagulation factor VII	1dan	406	6/107
Fibroblast growth factor 2	4fgf	146	6/18
Granulocyte colony-stimulating factor	1rhg	177	3/32
Interleukin-4	1rcb	129	5/18
Lysozyme C	1vfb	129	8/25
Neurotrophin-3	1nt3	119	4/33
T-cell surface antigen CD2	1cdc	99	6/14

Jelikož mnoho proteinů postrádá obecně používaný český název, jsou zde v zájmu sjednocení názvy všech uvedeny v angličtině. Poslední sloupec tabulky udává, kolik reziduí proteinu bylo zkoumáno pomocí ASM a kolik takových reziduí dosáhlo hodnot $\Delta\Delta G$ a $\Delta\Delta SASA$ pro zařazení mezi aktivní místa podle ASEdb [18]. Pro dříve používaný příklad interleukinu 4 bylo experimentálně potvrzeno pět aktivních míst z celkem osmnácti zkoumaných reziduí.

Tabulka 3. Vybrané proteiny - zdrojový organismus a příbuzné proteiny

PDB kód	Organismus	Příbuzné proteiny
1brs	Bacillus amyloliquefaciens	P39873, P79351, P87350, Q29542, Q29543
1dan	Homo sapiens	P22457, P70375, P98139, Q2F9P2, Q2F9P4, Q8K3U6
4fgf	Homo sapiens	P03969, P12226, P15655, P20003, P48798, P48800, Q5IS69, Q60487
1rhg	Homo sapiens	O02708, O02837, P09920, P35833, P35834, Q28746
1rcb	Homo sapiens	O77762, P07750, P30367, P30368, P42202, P79339, Q04745, Q8HYB1-2
1vfb	Gallus Gallus	P00702, P00703, P00704, P04421, P61626, P85345, Q7LZQ1
1nt3	Homo sapiens	P18280, P20181, P25433, P25435, Q06AV0, Q08DT3, Q9TST2
1cdc	Rattus norvegicus	P06729, P08920, P37998, Q6SZ61

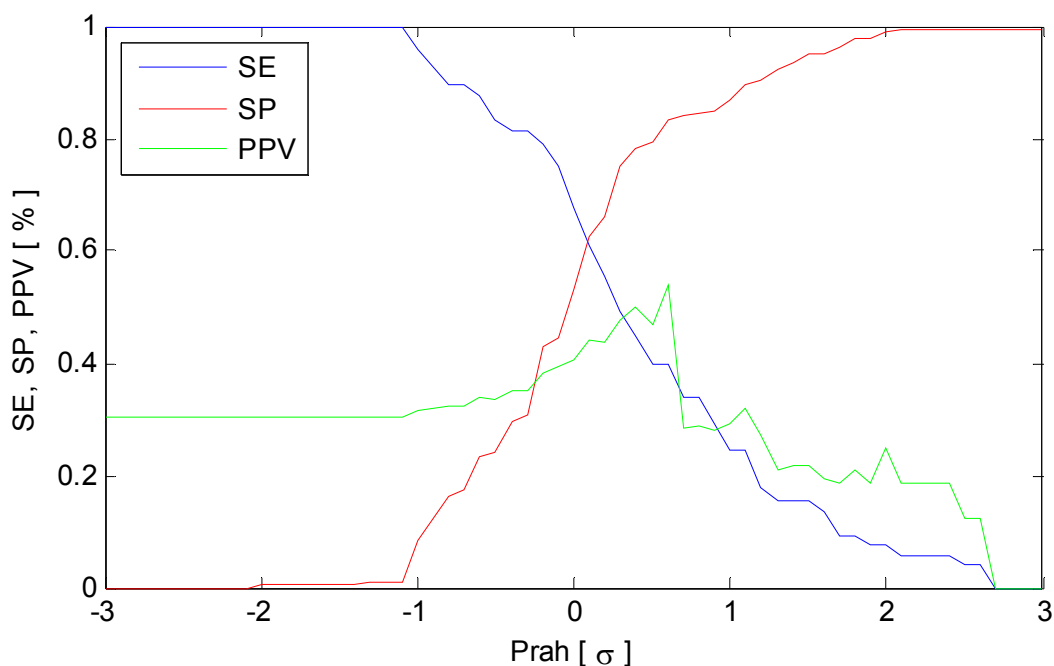
Jako aktivní místo je zde přitom označeno reziduum s $\Delta\Delta G \geq 1$ kcal/mol nebo $\Delta\Delta SASA \geq 100 \text{ \AA}^2$. Nutno podotknout, že hodnoty $\Delta\Delta SASA$ nejsou k dispozici u všech vybraných proteinů. U takových proteinů je potom jediným kritériem $\Delta\Delta G$.

Tabulka 3 shrnuje, z jakého organismu vybraný protein pochází, a které proteiny mu byly přiřazeny jako příbuzné. Pro každý zkoumaný protein bylo vybráno čtyři až osm proteinů příbuzných.

6.2 Metoda A

Jelikož způsob detekce aktivních míst v této metodě je založený na jednoduchém prahování, dosažené hodnoty statistických ukazatelů účinnosti jsou výší prahu samozřejmě ovlivněny. Jak se s volbou prahu mění, ukazuje obrázek 22.

Základní hodnota prahu je průměr signálu, na němž k prahování dochází, zde se prahem oproti této hodnotě pohybuje vždy o násobek směrodatné odchylky téhož signálu. Data, z nichž je graf vykreslen, jsou uvedena v tabulce 4. Jako popisné veličiny byly zvoleny senzitivita (SE), specificita (SP) a pozitivní prediktivní hodnota (*positive predictive value*, PPV). Jedná se přitom o průměrné hodnoty získané při zpracování osmi referenčních proteinů za použití příbuzných proteinů z tabulky 3.



Obrázek 22. Metoda A – závislost statistických ukazatelů na volbě prahu

Konkrétní hodnota prahu vždy závisí na uživateli a jeho potřebách. Pro přehledný popis algoritmu je však potřeba zvolit jednu reprezentativní hodnotu prahu a účinnost jaké při ní dosahuje. Standardně bych vybral bod, kde se protne specificita a senzitivita (kvůli diskrétní povaze vzorkování spíše nejbližší známý bod). V oblasti predikce aktivních míst je ale kladen důraz na co nejvyšší PPV [15], uvádím

proto navíc i hodnoty, při kterých algoritmus dosahuje nejvyšší možné PPV. Tyto údaje jsou přehledně uvedeny v tabulce 5. Zde se navíc objevuje i ukazatel AUC (*area under curve*) značící plochu pod ROC křivkou. ROC křivky jsou potom pro srovnání vykresleny na obrázku 32 v kapitole 7.

Tabulka 4. Statistické údaje o účinnosti metody A

Práh [σ]	SE [%]	SP [%]	PPV [%]	Práh [σ]	SE [%]	SP [%]	PPV [%]	Práh [σ]	SE [%]	SP [%]	PPV [%]
-2,1	100,00	0,00	30,42	-0,4	81,25	29,42	34,91	1,3	15,63	92,30	20,83
-2,0	100,00	0,66	30,52	-0,3	81,25	30,75	35,20	1,4	15,63	93,39	21,67
-1,9	100,00	0,66	30,52	-0,2	79,17	42,81	38,28	1,5	15,63	95,01	21,67
-1,8	100,00	0,66	30,52	-0,1	75,00	44,65	39,49	1,6	13,54	95,01	19,58
-1,7	100,00	0,66	30,52	0,0	67,71	53,32	40,70	1,7	9,38	96,40	18,75
-1,6	100,00	0,66	30,52	0,1	61,04	62,44	43,97	1,8	9,38	97,79	20,83
-1,5	100,00	0,66	30,52	0,2	55,31	66,12	43,67	1,9	7,81	97,79	18,75
-1,4	100,00	0,66	30,52	0,3	49,06	75,22	47,52	2,0	7,81	98,83	25,00
-1,3	100,00	0,91	30,53	0,4	44,90	78,26	50,12	2,1	5,73	99,26	18,75
-1,2	100,00	1,03	30,54	0,5	39,69	79,42	46,96	2,2	5,73	99,26	18,75
-1,1	100,00	1,15	30,54	0,6	39,69	83,47	54,11	2,3	5,73	99,26	18,75
-1,0	95,83	8,34	31,60	0,7	33,96	83,96	28,49	2,4	5,73	99,26	18,75
-0,9	92,71	12,44	31,96	0,8	33,96	84,33	28,96	2,5	4,17	99,26	12,50
-0,8	89,58	16,46	32,29	0,9	29,38	84,83	28,22	2,6	4,17	99,26	12,50
-0,7	89,58	17,38	32,42	1,0	24,38	86,76	29,32	2,7	0,00	99,26	0,00
-0,6	87,50	23,43	33,99	1,1	24,38	89,47	32,11				
-0,5	83,33	24,23	33,62	1,2	17,71	90,56	27,08				

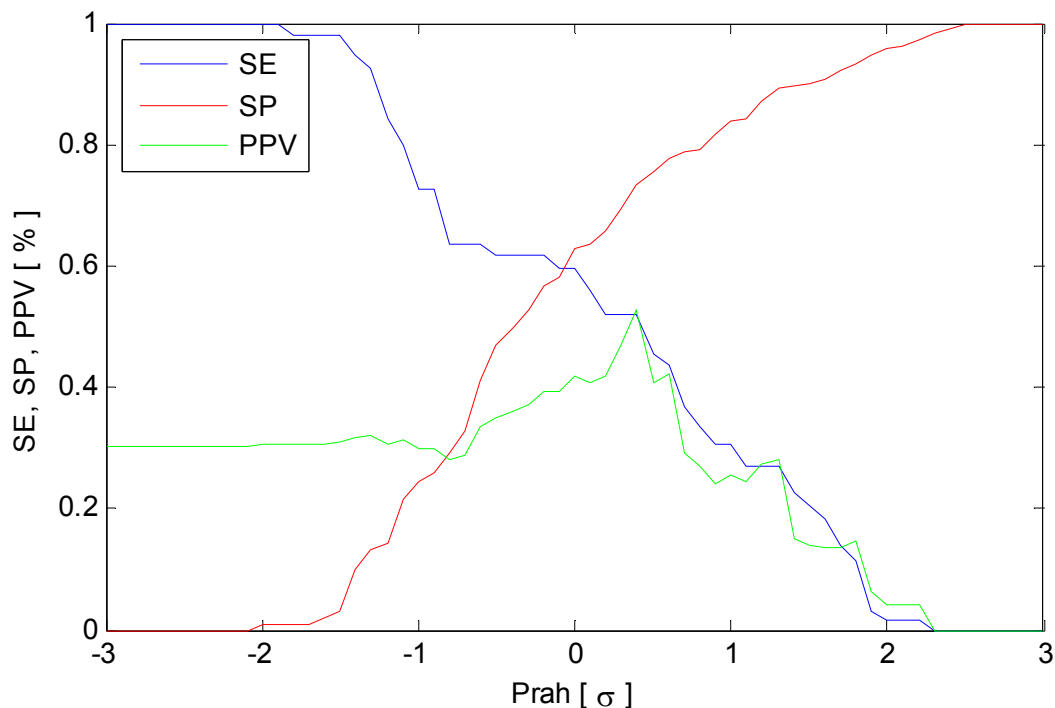
Tabulka 5. Metoda A – souhrnné ukazatele účinnosti

	Průsečík SE a SP	PPV_{max}
Práh [σ]	0,10	0,60
SE [%]	61,04	39,88
SP [%]	62,44	83,67
PPV [%]	43,97	54,11
AUC [%]	64,49	

6.3 Metoda B

Stejná data, kterými byla v předchozí kapitole popsána metoda A, jsou zde uvedena i pro metodu B. Rozsah hodnot prahu v tabulce 6 se oproti tabulce 4 liší z toho důvodu, že další kroky daným směrem by nepřinesly žádné nové informace. Jednalo by se jen o opakování řádků s jedinou změnou v hodnotě prahu. Vykreslení

charakteristik na obrázku 23 je ovšem provedeno v plném rozsahu. Souhrnné charakteristiky jsou potom uvedeny v tabulce 7.



Obrázek 23. Metoda B – závislost statistických ukazatelů na volbě prahu

Tabulka 6. Statistické údaje o účinnosti metody B

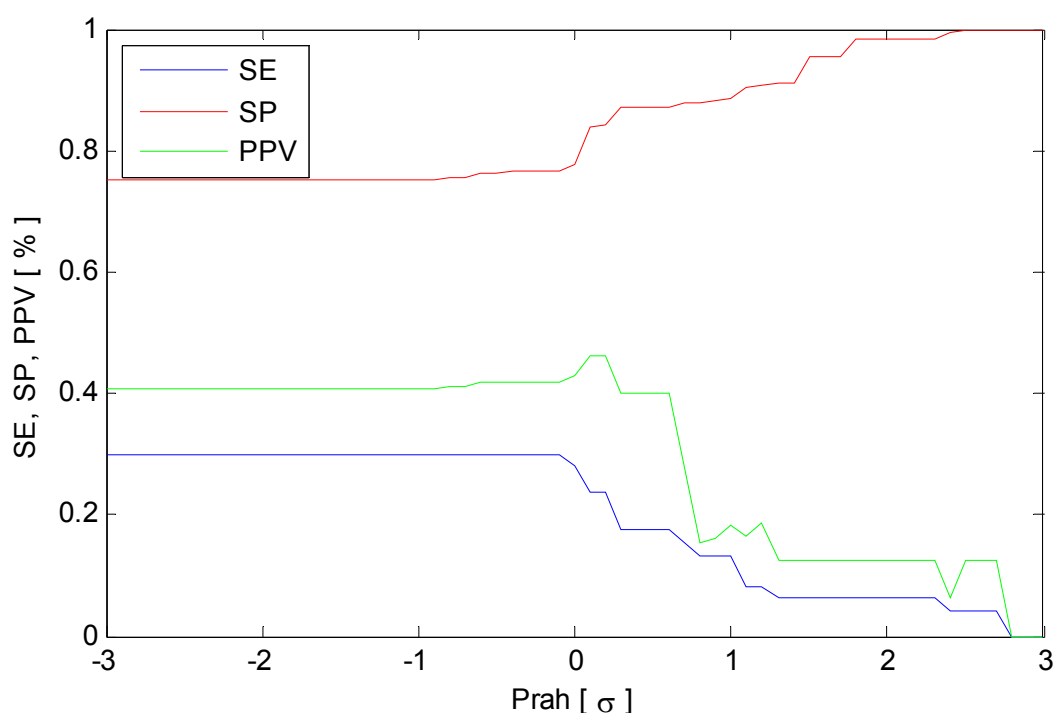
Práh [σ]	SE [%]	SP [%]	PPV [%]	Práh [σ]	SE [%]	SP [%]	PPV [%]	Práh [σ]	SE [%]	SP [%]	PPV [%]
-2,1	100,00	0,00	30,42	-0,5	61,67	47,04	34,99	1,1	26,88	84,36	24,63
-2,0	100,00	0,74	30,58	-0,4	61,67	49,96	36,09	1,2	26,88	87,04	27,26
-1,9	100,00	0,74	30,58	-0,3	61,67	52,66	37,05	1,3	26,88	89,28	27,96
-1,8	97,92	0,74	30,47	-0,2	61,67	56,83	39,31	1,4	22,71	89,53	14,88
-1,7	97,92	0,86	30,48	-0,1	59,58	58,15	39,13	1,5	20,63	89,90	13,99
-1,6	97,92	1,94	30,69	0,0	59,58	62,88	41,74	1,6	18,13	90,88	13,69
-1,5	97,92	3,16	30,83	0,1	55,94	63,56	40,68	1,7	13,96	92,28	13,75
-1,4	94,79	9,96	31,73	0,2	51,88	65,59	41,75	1,8	11,46	93,39	14,58
-1,3	92,71	13,12	31,93	0,3	51,88	69,50	46,94	1,9	3,13	94,90	6,25
-1,2	84,38	14,33	30,50	0,4	51,88	73,41	52,55	2,0	1,56	95,99	4,17
-1,1	79,79	21,67	31,40	0,5	45,63	75,44	40,69	2,1	1,56	96,11	4,17
-1,0	72,50	24,59	29,82	0,6	43,54	77,76	42,24	2,2	1,56	97,20	4,17
-0,9	72,50	26,03	30,00	0,7	36,77	78,91	29,13	2,3	0,00	98,28	0,00
-0,8	63,75	29,08	28,02	0,8	33,65	79,16	27,10	2,4	0,00	99,14	0,00
-0,7	63,75	32,81	28,78	0,9	30,52	81,71	24,06	2,5	0,00	100,00	0,00
-0,6	63,75	41,13	33,43	1,0	30,52	83,74	25,63				

Tabulka 7. Metoda B – souhrnné ukazatele účinnosti

	Průsečík SE a SP	PPV _{max}
Práh [σ]	-0,10	0,40
SE [%]	59,58	51,88
SP [%]	58,15	73,41
PPV [%]	39,13	52,55
AUC [%]	58,02	

6.4 Metoda C

Na rozdíl od předchozích metod, metoda C nepoužívá prahování, ale detekci okolí peaku popsanou v kapitole 4.3. Důsledkem je nestandardní průběh statistických parametrů SE, SP a PPV, ten je znázorněn na obrázku 24. Vykreslené průběhy reprezentují průměrné chování algoritmu při nasazení na všechny referenční proteiny za šířky detekce 2.

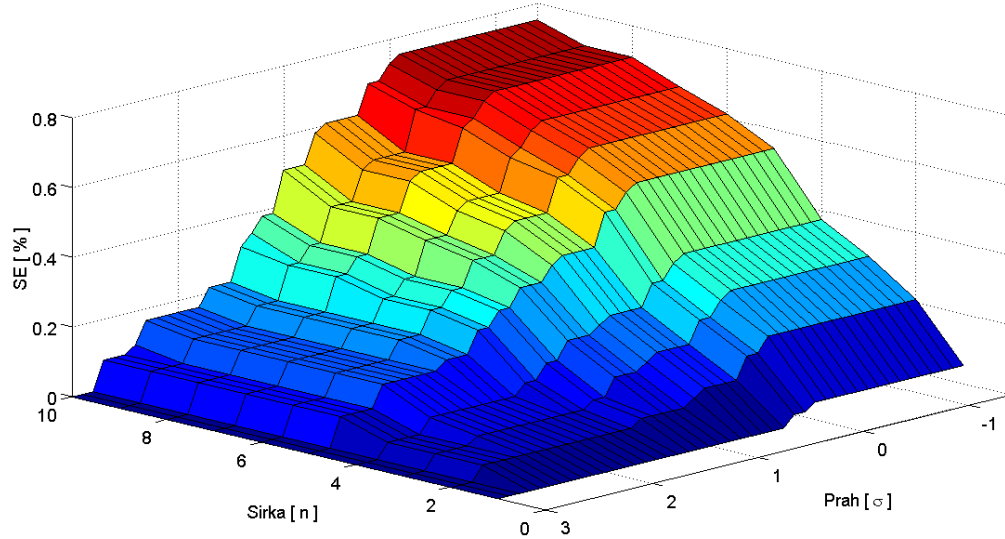


Obrázek 24. Metoda C – závislost statistických ukazatelů na volbě prahu

Hodnoty senzitivity a specificity se sice stále pohybují mezi jedničkou a nulou, pouhou volbou prahu ale nikdy nedosáhneme obou extrémů. Ze zde uvedených průběhů by se mohlo zdát, že jde jen o pravou část grafu, není tomu ale tak.

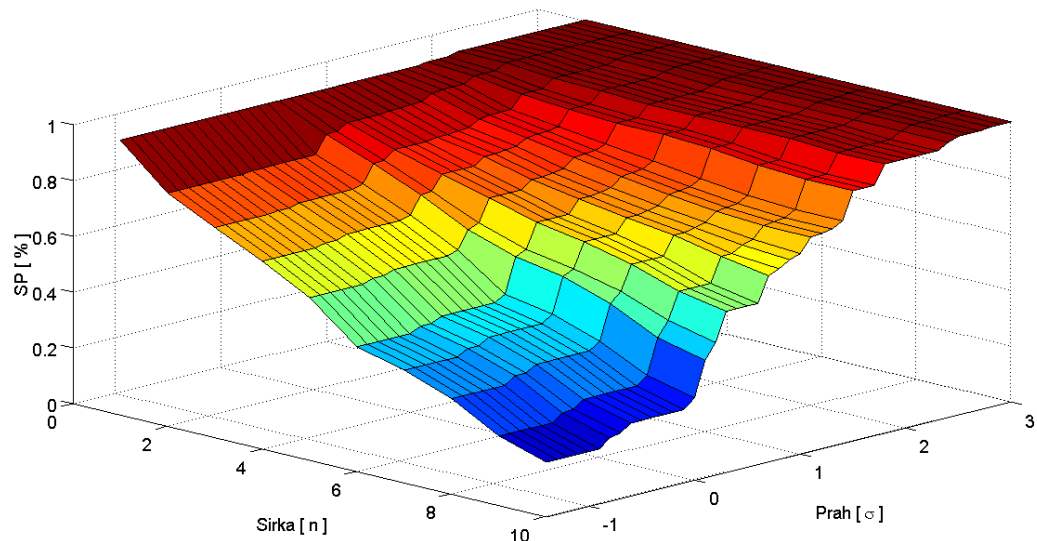
Od jistého prahu (zde zhruba 1) algoritmus detekuje stále stejné vzorky, ať už klesáme sebevíc. To je způsobeno tím, že práh už je pod úrovní nejnižších peaků. Všechny vzorky v okolí vrcholů jsou tím pádem detekovány. Tato oblast se další manipulací s prahem nerozšíří, k tomu by bylo potřeba změnit šířku detekce,

neboli velikost okolí peaků, ve kterém jsou vzorky považovány za aktivní místa. Nelze proto nijak jednoduše vykreslit ROC křivku, ani určit, jaká je pod ní plocha.



Obrázek 25. Metoda C – závislost senzitivity na volbě prahu a šířky detekce

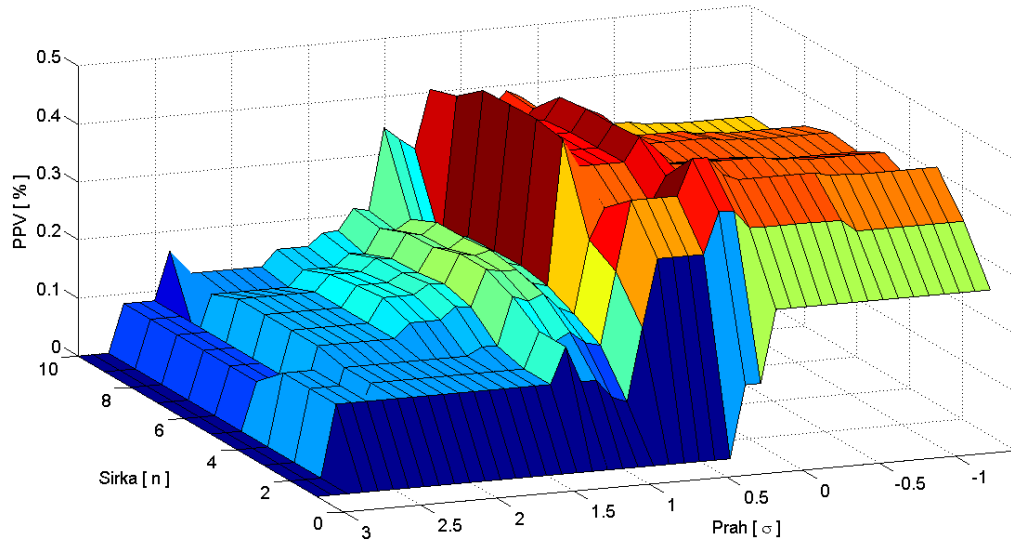
Účinnost algoritmu tedy závisí na dvou parametrech: velikosti prahu a šířce detekce. Obrázky 25 až 27 ukazují tuto závislost pokaždé pro jeden z ukazatelů SE, SP, nebo PPV. Vykreslovaná oblast je zde zdola omezena prahem -1,4, protože – jak už bylo popsáno – u nižších prahů už se tyto parametry dále nemění.



Obrázek 26. Metoda C – závislost specificity na volbě prahu a šířky detekce

Obecně můžeme říct, že senzitivita s šířkou detekce roste, zatímco s velikostí prahu klesá. Specificita se dle očekávání chová přesně naopak; s šířkou detekce klesá

a s velikostí prahu stoupá. Toto chování popisují obrázky 25 a 26. Při jejich prohlížení je třeba dát si pozor na orientaci os, ta je totiž v zájmu lepší viditelnosti průběhu ukazatele u každého z nich odlišná.



Obrázek 27. Metoda C – závislost pozitivní prediktivní hodnoty na volbě prahu a šířky detekce

Průběh pozitivní prediktivní hodnoty vykreslený na obrázku 27 už není tak jednoznačný. Sice zde platí, že s nižší velikostí prahu dosahuje algoritmus vyšší hodnoty PPV, mezi hodnotami prahu nula a jedna se ale nachází jakýsi dvojité hřeben, kde je možno dosáhnout vyšších hodnot PPV než kdekoli jinde. Závislost na šířce detekce je ještě méně zřejmá, nejlépe proto bude držet se chování zmiňovaného hřebene. Ten při volbě šířky detekce 3 dosahuje peaku představujícího globální maximum. Další plochý extrém můžeme pozorovat pro šířku detekce 5 – 8.

Tabulka 8. Metoda C – souhrnné ukazatele účinnosti

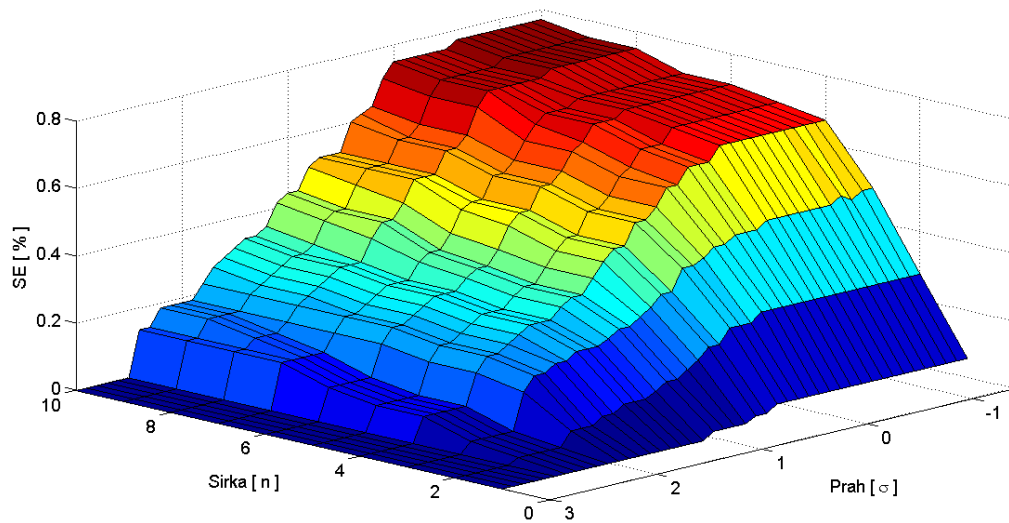
Nastavení	SEP	PPV _{max}
Práh [σ]	0,70	0,20
Šířka detekce [n]	10	3
SE [%]	58,33	23,75
SP [%]	66,56	84,13
PPV [%]	40,46	46,13

Jelikož statistické ukazatele závisí na dvou nezávislých parametrech, průsečík průběhu specificity a senzitivity netvoří bod, ale křivka. Pro přehledové zhodnocení metody proto volím nepatrně odlišný ukazatel, než u metod předchozích. Místo hodnot v průsečíku uvádím hodnoty ukazatelů v bodě, kde dochází k maximalizaci jednoho ze dvou parametrů: součtu SP + SE (nastavení SEP), nebo PPV (nastavení PPV_{max}).

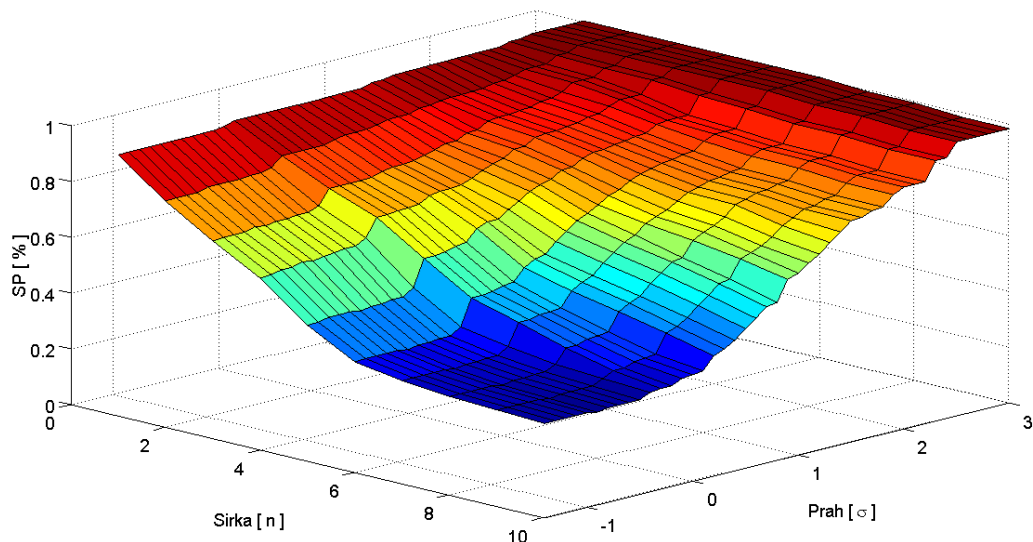
Tyto údaje jsou uvedeny v tabulce 8. Podle hodnot prahu a šířky detekce je vidět, že se algoritmus v obou případech pohybuje v oblasti výše popsaného hřebene.

6.5 Metoda D

Metoda D je svou podstatou velice podobná metodě C, z toho důvodu byl zvolen i stejný způsob hodnocení. Závislost senzitivity, specifity a pozitivní prediktivní hodnoty na volbě hodnoty prahu a šířky detekce je vykreslena na následujících obrázcích.

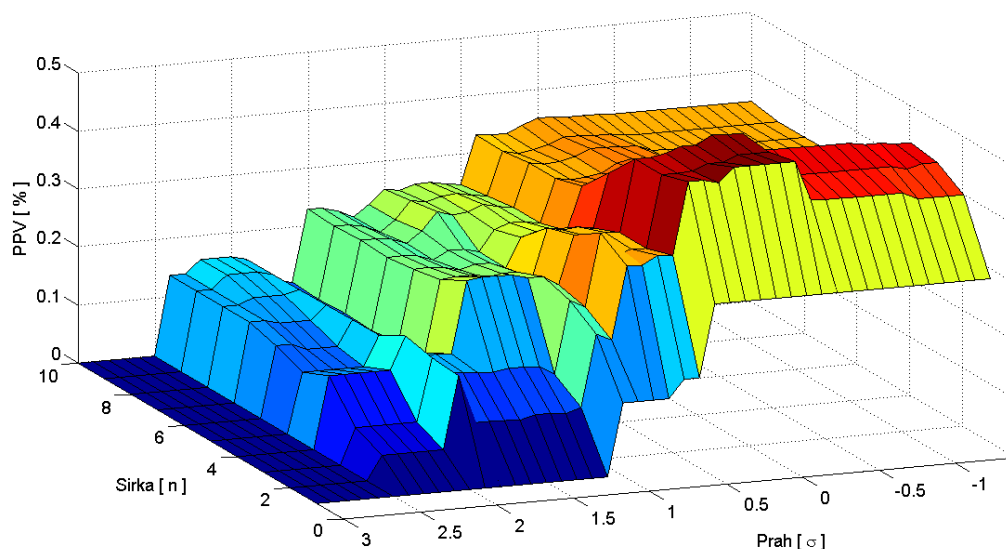


Obrázek 28. Metoda D – závislost senzitivity na volbě prahu a šířky detekce



Obrázek 29. Metoda D – závislost specifity na volbě prahu a šířky detekce

Průběh senzitivity a specifity je zde velmi podobný. Senzitivita opět roste s šířkou detekce a naopak klesá s hodnotou prahu. Specifita se chová přesně naopak. To je možno vidět na obrázcích 28 a 29.



Obrázek 30. Metoda D – závislost pozitivní prediktivní hodnoty na volbě prahu a šířky detekce

Průběh PPV je však mírně odlišný. S rostoucím prahem PPV obecně klesá. Na intervalu $(-0,5; 0,5)$ ale máme plochý peak obsahující nejvyšší hodnoty v celé vykreslované oblasti. Z hlediska šířky detekce se tento peak nachází mezi hodnotami 2 a 5. Tato oblast šířky detekce zároveň dosahuje nadprůměrných hodnot PPV i při většině jiných hodnot prahu.

Když se podíváme na souhrnné statistické ukazatele v tabulce 9, vybrané stejným způsobem jako u metody C, vidíme, že v obou případech se pohybujeme právě v oblasti tohoto plochého peaku.

Tabulka 9. Metoda D – souhrnné ukazatele účinnosti

Nastavení	SEP	PPV _{max}
Práh [σ]	-0,1	-0,3
Šířka detekce [n]	4	2
SE [%]	62,50	27,50
SP [%]	73,32	86,80
PPV [%]	47,94	48,27

7 Diskuze

V této práci byly představeny dva způsoby detekce aktivních míst a dva odlišné signály, na nichž je lze detekovat. Byly zde popsány metody reprezentující všechny kombinace těchto postupů. V tomto ohledu jejich rozdělení přehledně ukazuje tabulka 10, kde řádky reprezentují způsob detekce a sloupce signál, na němž detekce probíhá.

Tabulka 10. Kategorizace navrhovaných metod

		Použitý signál	
		Výstup IST	Řez ST spektra
Způsob detekce	Prahování	A	B
	Detekce	C	D

7.1 Srovnání navrhovaných metod

Tabulka 11 přehledně shrnuje hodnoty statistických ukazatelů, kterých všechny metody dosáhly. Metody porovnávám ve dvou nastaveních. Jedním je PPV_{max} , což je nastavení maximalizující PPV. Druhým je SEP , které slučuje nastavení použité u metod A a B k dosažení průsečíku jejich hodnot SE a SP, a nastavení maximalizující u metod C a D součet SE+SP. Parametr AUC na nastavení sice závislý není, pro úplnost ho zde ale uvádím také. Průběh ROC křivek pro metody A a B je vykreslen na obrázku 32. Některé z uvedených ukazatelů jsou potom vyneseny v grafu na obrázku 31.

Tabulka 11. Srovnání přehledových ukazatelů navrhovaných metod

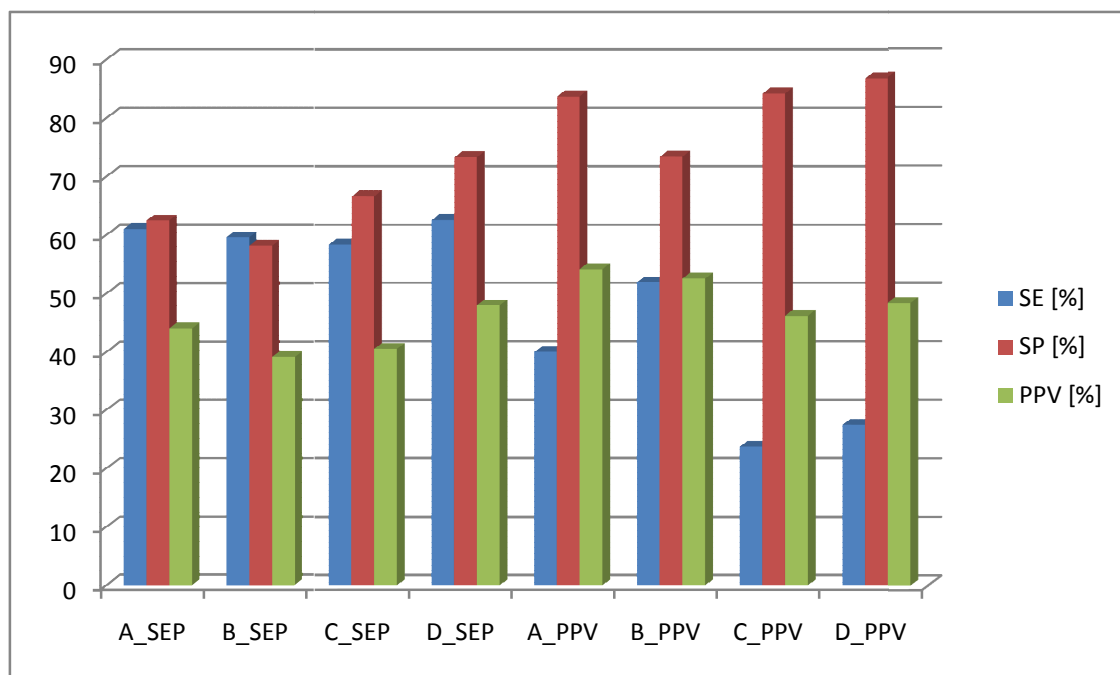
Nastavení	SEP				PPV _{max}			
Metoda	A	B	C	D	A	B	C	D
Šířka detekce [n]	-	-	10	4	-	-	3	2
Práh [σ]	0,1	-0,1	0,7	-0,1	0,6	0,4	0,2	-0,3
SE [%]	61,04	59,58	58,33	62,5	39,88	51,88	23,75	27,5
SP [%]	62,44	58,15	66,56	73,32	83,67	73,41	84,13	86,8
PPV [%]	43,97	39,13	40,46	47,94	54,11	52,55	46,13	48,27
AUC [%]	64,49	58,02	-	-	64,49	58,02	-	-

7.1.1 Nastavení SEP

V části tabulky 11 pro nastavení SEP vidíme, že nejlepších výsledků dosahuje metoda D. Za druhou nejlepší metodu by se dala označit metoda A, která má jen zhruba o procento menší SE. V ostatních ukazatelích potom zaostává o poznání výrazněji. U metody D bohužel není k dispozici parametr AUC, ten tedy srovnat nelze.

Tento výsledek je velice zajímavý, metody A a D totiž tvoří jednu z nejvíce rozdílných dvojic ze všech algoritmů. Každá používá k detekci aktivních míst odlišný signál (filtrovaný a upravený výstup IST pro A; řez ST spektrem pro D) a dokonce

i odlišný způsob detekce (jednoduché prahování pro A; detekce vrcholů peaků a jejich sousedních vzorků pro D). Podobné jsou si snad jen prahy, které obě metody k dosažení tohoto výsledku používají, oba se drží velice blízko nuly.



Obrázek 31. Srovnání statistických ukazatelů

7.1.2 Nastavení PPV_{max}

Nalezení nejlepší metody v nastavení PPV_{max} už není tak jednoduché. Jelikož cílem tohoto nastavení je maximalizace PPV, dalo by se uvažovat o metodě A. Ta sice dosahuje nejvyššího PPV, ve SP ji ale předčí metody C a D – byť jen o jednotky procent – v SE zase metoda B, a to velice výrazně. Metoda B při tomto nastavení dosahuje nejvyrovnanějších výsledků. Oproti ostatním sice zaostává ve SP, má ale bezkonkurenčně nejvyšší SE a relativně vysokou PPV.

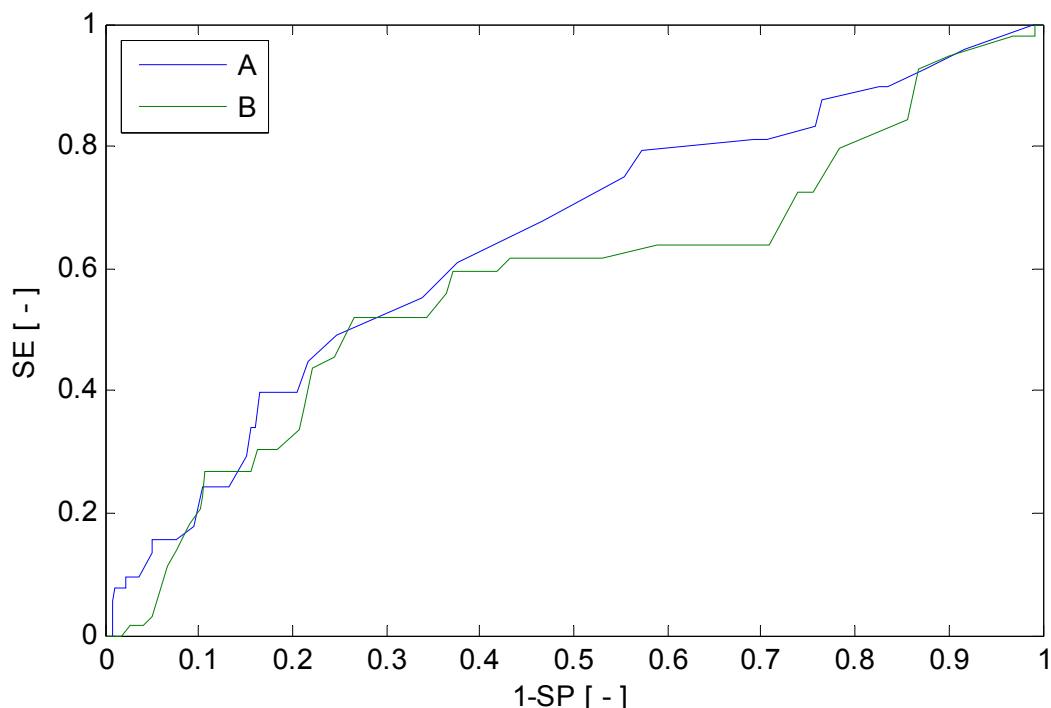
V takové situaci si netroufám označit některou z metod jako nejlepší. Pokud by si někdo měl vybrat, kterou z nich použít, měl by se řídit vlastními potřebami, případně vzít v úvahu výsledky více metod.

7.1.3 SEP versus PPV_{max}

Porovnáme-li obě nastavení mezi sebou, můžeme obecně říct, že při nastavení SEP dosahujeme ve všech případech vyšší SE. Při nastavení PPV_{max} je potom možno dosáhnout jak vyšší SP, tak PPV. Zatímco pokles SE a růst SP je velmi výrazný, o růstu PPV už se to ve všech případech tvrdit nedá.

Při pohledu na výsledky metody D dosažené při obou nastaveních by se dalo říct, že nastavení PPV_{max} přináší příliš velký pokles SE, aby to bylo vykoupeno růstem SP a PPV. SE zde oproti nastavení SEP klesá o celých 35 %, zatímco zisk je pouhých

0,33 % PPV a 13,48 % SP. Na první pohled tedy oproti nastavení SEP nejde o krok správným směrem. Věřím, že v určitých případech však toto nastavení přesto může být žádoucí.



Obrázek 32. ROC křivka pro metody A a B

7.2 Použití serveru Hotpoint jako reference

V předchozí kapitole bylo ukázáno, jakých výsledků navrhované metody dosahují při použití referenčních dat z ASEdb, tedy dat, získaných experimentální cestou. Situace na poli predikce aktivních míst je ovšem taková, že téměř každý vyvinutý výpočetní nástroj si kolem sebe vytvoří separátní databázi založenou na svých vlastních výstupech. Uvádím zde proto i výsledky navrhovaných metod při použití referenčních dat právě z jedné takové databáze, Hotpointu [35]. Ty jsou uvedeny v tabulce 12 za použití stejného postupu jako v tabulce 11.

Tabulka 12. Srovnání souhrnných ukazatelů za použití referenčních dat ze serveru Hotpoint

Nastavení	SEP				PPV _{max}			
	A	B	C	D	A	B	C	D
Metoda								
Šířka detekce [n]	-	-	10	8	-	-	1	1
Práh [σ]	-0,1	-0,5	0,7	1,3	0,5	1,5	0,2	1,1
SE [%]	50,92	50,72	50,19	24,07	26,63	10,25	4,67	3,03
SP [%]	47,70	51,68	57,19	75,29	73,77	92,61	97,17	98,59
PPV [%]	23,93	25,78	23,97	14,24	27,45	32,91	37,44	31,40
AUC [%]	50,64	52,2	-	-	50,64	52,2	-	-

Je zde vidět, že za použití těchto referencí dosahují všechny metody výrazně horších výsledků. Nejvíce utrpěl ukazatel PPV, který teď místo dřívějších 40 – 55 % dosahuje sotva 30 %. Tuto hranici výrazněji překročila pouze metoda C za nastavení PPV_{max} , nejvýraznější pokles je naopak možno pozorovat u metody D za nastavení SEP.

Obecně lze říct, že při nastavení PPV_{max} bylo oproti nastavení SEP dosahováno vyšších hodnot PPV (s tím cílem je toto nastavení také koncipováno), ovšem za cenu výrazně nižší SE. Vedlejším efektem je potom výrazný růst SP, tato skutečnost zde ale nepředznamenává nic pozitivního. Pokles SE a růst SP je zde přitom mnohem výraznější, než při referenci oproti ASEdb.

Oproti navrhovaným metodám však Hotpoint hledá aktivní místa vždy pro konkrétní pár interagujících proteinů. Jako protějšek byl proto zadán stejný protein, údaje poskytnuté Hotpointem tedy platily pro reakce, při níž vzniká homodimer. Navrhované metody informaci o proteinu, s nímž má interakce probíhat, nevyužívají. Tabulka 13 pro úplnost uvádí, které řetězce byly z hlediska Hotpointu použity.

Tabulka 13. Řetězce referenčních proteinů pro srovnání s Hotpointem

Název proteinu	Řetězec
Barnase	A
Coagulation factor VII	L
Fibroblast growth factor 2	A
Granulocyte colony-stimulating factor	A
Interleukin-4	A
Lysozyme C	C
Neurotrophin-3	A
T-cell surface antigen CD2	A

7.3 Vliv volby příbuzných proteinů

Volba příbuzných proteinů může výsledek analýzy ovlivnit zcela zásadním způsobem, přitom se nepodařilo najít žádná pravidla odhalující jak tuto volbu provádět a dosáhnout uspokojivého výsledku. Je dokonce možné, že odlišnou volbou proteinů příbuzných k referenčním by bylo možno dosáhnout lepších výsledků, než zde uvádím, možností je ale příliš mnoho.

Špatná volba příbuzných proteinů se může projevit dvěma způsoby: charakteristická frekvence zůstane zachována, dojde ale ke změně účinnosti; nebo charakteristická frekvence zachována nezůstane a změna účinnosti pak zpravidla bude o to větší.

7.3.1 Zachování charakteristické frekvence

Jako příklad nám zde opět poslouží interleukin 4, v tabulce 14 jsou uvedeny původní hodnoty účinnosti dosahované na tomto proteinu a hodnoty po odebrání

příbuzného proteinu P30367. Údajů bylo dosaženo s prahem rovným průměru signálu (tedy s nulovým posunutím) a u metody C s šířkou detekce 3.

Vidíme, že vypuštěním sekvence P30367 z analýzy interleukinu 4 došlo u metody A k propadu všech tří ukazatelů. Nejvýrazněji přitom byla postižena senzitivita. Metoda C se v tomto ohledu ukázala robustnější. Její senzitivita zůstala zachována a ani pokles zbývajících dvou ukazatelů nebyl tak výrazný, přesto i zde dostáváme jednoznačně horší výsledky.

Tabulka 14. Změna účinnosti při odebrání P30367

	Původní			Bez P30367		
	Char. frekvence = 0,59			Char. frekvence = 0,59		
	SE [%]	SP [%]	PPV [%]	SE [%]	SP [%]	PPV [%]
Metoda A	100,00	30,77	35,71	60,00	23,08	23,08
Metoda C	40,00	53,85	25,00	40,00	46,15	22,22

Metody B a D zde neuvádím, ty totiž aktivní místa predikují z průběhu charakteristické frekvence v ST spektru. Jelikož charakteristická frekvence byla zachována, nedochází u nich k žádné změně.

7.3.2 Změna charakteristické frekvence

I tento případ budu ilustrovat na interleukinu 4, tentokrát ovšem vypustím protein P30368. Kromě toho jsou všechny parametry zachovány stejně jako v předchozí kapitole. Výsledky jsou uvedeny v tabulce 15.

Tabulka 15. Změna účinnosti při odebrání P30368

	Původní			Bez P30368		
	Char. frekvence = 0,59			Char. frekvence = 0,72		
	SE [%]	SP [%]	PPV [%]	SE [%]	SP [%]	PPV [%]
Metoda A	100,00	30,77	35,71	0,00	53,85	0,00
Metoda B	60,00	30,77	35,71	80,00	38,46	33,33
Metoda C	40,00	53,85	25,00	0,00	84,62	0,00
Metoda D	80,00	46,15	36,36	40,00	23,08	16,67

Změna charakteristické frekvence se už odrazí na všech navrhovaných metodách. Vidíme, že nejhůře zde dopadly metody A a C, které nyní nejsou schopné detekovat jediné aktivní místo správně. Výrazný pokles nastal i u metody D, kde se všechny ukazatele dostaly na polovinu svých původních hodnot. Pouze metoda B zaznamenala jisté zlepšení, což lze pravděpodobně přičíst vlivu náhody. Obecně je ale zřejmé, že při změně charakteristické frekvence dojde k výrazné změně účinnosti. V případě, že původní charakteristická frekvence byla špatně zvolená, může být změna k dobrému, pro tuto možnost zde ale příklad uvádět nebudu.

Nutno poznamenat, že zatímco změny specifity a pozitivního prediktivního faktoru jsou relativně plynulé, změny senzitivity se odehrávají vždy ve skocích o nejméně 20 %. To je způsobeno tím, že v sekvenci interleukinu 4 se nachází pouze pět experimentálně potvrzených aktivních míst, menší skoky zde proto možné nejsou.

7.4 Srovnání s výsledky jiných skupin

Tabulka 16 ukazuje přehledné srovnání navrhovaných metod s metodami jiných skupin. Projevuje se zde nešvar, kterým je tato oblast často postihována – k obecnému srovnání je k dispozici velmi málo statistických údajů. Některé skupiny ve svých publikacích jednoduše neuvádějí úspěšnost svých metod, nebo uvádějí jiné ukazatele než ostatní. Do tohoto srovnání na rozdíl od těch předchozích proto uvádím pouze dva nejčastěji uváděné ukazatele, a to senzitivitu a pozitivní prediktivní hodnotu.

Vidíme, že hodnoty senzitivity navrhovaných metod jsou často srovnatelné s jinými metodami, nutno však uznat, že téměř vždy nižší. Tomu se vymyká pouze metoda označená jako *ST filtering*, ta dosahuje výrazně vyšší senzitivity než všechny uvedené metody. *ISIS* naopak o poznání nižší. Pro pořádek uvádím, že jde o metodu popsanou v kapitole 3.3.3. Co se PPV týče, tam všechny navrhované metody viditelně zaostávají. Ani při jeho maximalizaci v tomto ukazateli nedosahují takových hodnot jako jiné skupiny.

Tabulka 16. Srovnání navrhovaných metod s jinými metodami. Údaje částečně převzaty z [9], [25]

		SE [%]	PPV [%]
ISIS		15	89
ROBETTA		69	71
MAPPIS		66	63
HotPoint		59	70
pyDockNIP		43	75
ST filtering		83,33	62,5
SEP	A	61,04	43,97
	B	59,58	39,13
	C	58,33	40,46
	D	62,5	47,94
PPV_{max}	A	39,88	54,11
	B	51,88	52,55
	C	23,75	46,13
	D	27,5	48,27

8 Popis přiloženého softwaru

Zde popsaný algoritmus je k této práci i přiložen. Implementován byl v programovém prostředí MATLAB verze 7.11.0.584 (R2010b). Toto je zároveň jediná verze MATLABu, na které byla testována jeho funkčnost. Soubory každé z metod se nacházejí ve stejnojmenné složce a spouštěcí funkcí je vždy START.

8.1 Vstupní data

Algoritmus na svém vstupu vyžaduje složku obsahující sekvence proteinů. Zkoumaný protein musí být vždy soubor, jehož název je abecedně první. Zároveň by se mezi jemu příbuznými proteiny neměla vyskytovat kratší sekvence. Pokud se tak stane, analýza sice proběhne, uživatel ale bude upozorněn, že na takové výsledky se nelze spoléhat. Zárukou správného výpočtu je používat pro zkoumaný protein identifikátor PBD a pro příbuzné proteiny identifikátor Uniprot.

8.2 Volání funkce

Uživatel může cestu ke složce se sekvencemi proteinů specifikovat při volání funkce pomocí příkazového řádku, v takovém případě bude cesta zapsána do posledního vstupu jako textový řetězec. Pokud toto uživatel neprovede, objeví se po spuštění funkce okno, ve kterém bude moct složku zvolit ve stromové struktuře.

Pro metody A a B je tedy povinný jeden vstup a to hodnota prahu uvedená jako násobek směrodatné odchylky signálu, který bude následně přičten k jeho průměrné hodnotě. Metody C a D potřebují navíc zadat i šířku detekce. Za těmito vstupy potom může následovat již zmiňovaná cesta ke složce se sekvencemi.

8.3 Výstup analýzy

Výsledek analýzy je zobrazen v příkazovém řádku spolu s identifikací zpracovaného proteinu a souboru, z něhož byla jeho sekvence načtena (uživatel tak má jistotu, že vstupní data poskytl ve správném formátu). Kromě toho jsou pozice nalezených aktivních míst zapsány do textového souboru se stejným názvem, jako má složka se sekvencemi. Tento soubor je uložen v aktivní složce MATLABu v podsložce *Vystup*. Pokud zde již stejnojmenný soubor existuje, je automaticky přepsán.

Ukládání výstupu do souboru nese za následek skutečnost, že při spuštění z CD nebo jiného read-only media algoritmus vypíše chybu. V takovém případě analýza sice proběhne, jediným výstupem je ale výpis v příkazovém řádku, případně ve výstupní proměnné funkce.

Závěr

Tato práce shrnuje výsledky některých dnešních metod určených pro výpočetní predikci aktivních míst v proteinech. Situace na tomto poli se už desítky let nemůže hnout z místa, kdy laboratorní metody jsou drahé a časově náročné, zatímco výpočetní metody jsou pořád ještě nespolehlivé. Vývoj přesto nestagne a je zde vidět posun vpřed. Stále se objevují přístupy, které ve své účinnosti předčí ty předchozí. Dá se proto očekávat, že jednou zde skutečně bude algoritmus schopný aktivní místa s dostatečnou přesností predikovat.

Navrhovaný algoritmus je zde prezentován ve čtyřech podobách, jejichž odlišnosti shrnuje tabulka 10. Všechny z nich přitom vycházejí jen ze sekvence proteinů. Není zde použita žádná strukturální informace. Každá metoda logicky dosahuje jiných výsledků, a i když je můžeme ze dvou různých pohledů rozdělit do dvou samostatných kategorií, nedá se říct, že by některý způsob detekce nebo modulace signálu vykazoval výrazně lepší výsledky než jiný. Všechny však alespoň za určitého nastavení dosahují relativně dobré účinnosti a za postupy jiných skupin příliš nezaostávají. Oproti nim dosahují navrhované metody srovnatelné senzitivity, pokulhávají ale v pozitivní prediktivní hodnotě. Výhodou může být jednoduchost navrhovaného postupu a relativní výpočetní nenáročnost, kdy většinu výpočetního času zabere pouze výpočet S-transformace, všechny další kroky potom představují minimální zátěž procesoru. Dá se tedy očekávat, že s nasazením algoritmu pro rychlý výpočet S-transformace (obdoba *fast fourier transform*), by se zátěž procesoru mohla ještě snížit.

Průběh signálů, na nichž jsou aktivní místa detekována, se ale zdá příliš pozvolný a hladký. Vzhledem k tomu, jak ojedinělá mají aktivní místa v sekvencích proteinů být, člověk by zde naopak čekal ostré peaky. Ty zde ale nevidíme. Příčinou může být skutečnost, že aktivní místa nejsou jedinými konzervovanými pozicemi v sekvenci. Jak bylo popsáno v kapitole 1.2, aktivní místa jsou obklopena rezidui tvořícími *O-ring*, která jsou také – byť ne v takové míře – evolučně konzervativní. Možná jsou to právě tato rezidua, která algoritmus detekuje jako falešně pozitivní.

Obecně je problém v tom, že experimentálních údajů o jednotlivých proteinech je velice málo. U zde vybraných referenčních proteinů šlo většinou o 20 % všech rezidui v sekvenci proteinů, přitom zhruba jednu pětina z těchto míst lze označit jako aktivní místo. Ve výsledku je tedy účinnost jakéhokoli algoritmu dána pouze těmito rezidui. Určitě by bylo vhodné financovat další experimenty, které by pomocí ASM sestavily skutečnou referenční databázi, třeba jen velmi omezeného množství proteinů, na nichž by potom mohly být výpočetní metody testovány. Tím by byl vyřešen i ten problém, že dnes je každá metoda testována na jiném vzorku proteinů, čímž jejich

vzájemná srovnatelnost značně trpí. Je ale zřejmé, co podobné snahy brzdí. Nejúspěšnější výpočetní metody dnes dosahují účinnosti více než 70 % a stále se objevují lepší a lepší. Za takové situace je otázkou, jestli by se podobná investice skutečně vyplatila.

Nadějným krokem by ovšem bylo zpracovat výsledky všech výpočetních metod do jedné databáze. Společným konsensem by tak mohlo vyjít najevo, že aktivní místa jsou dostatečně přesně predikována už dnes. Dnešní roztržitěné databáze, které si každá metoda tvoří kolem sebe, mají každá zvlášť jen omezenou použitelnost.

V kapitole 2.3 bylo ukázáno, že dvě hlavní kritéria pro klasifikaci reziduí jako aktivních míst – $\Delta\Delta G$ a $\Delta\Delta SASA$ – spolu jen špatně korelují. Tato skutečnost odporuje všem teoretickým předpokladům a stejně tak i selskému rozumu. Dosavadní teorie zatím nejsou schopné objasnit, proč tomu tak je. Je proto možné, že se už v základních teoriích nachází skrytá chyba, nebo v nich naopak chybí důležitý poznatek, který by tento nesoulad uvedl na pravou míru a otevřel tak cestu ke skutečné predikci aktivních míst.

Seznam literatury

- [1] RESEARCH COLLABORATORY FOR STRUCTURAL BIOINFORMATICS. *The Protein Data Bank* [online]. 1998, 2012-11-27 [cit. 2012-12-02]. Dostupné z: <http://www.rcsb.org/pdb>
- [2] MURRAY, Robert K. *Harperova biochemie*. 23. vyd. Jinočany: H+H, 2002, ix, [3], 872 s. ISBN 80-731-9013-3.
- [3] ZHOU, Zhaolan et al. Comprehensive proteomic analysis of the human spliceosome. *Nature*. 2002-9-12, vol. 419, no. 6903, s. 182 – 185. ISSN 0028-0836. DOI: 10.1038/nature01031. Dostupné z: <http://www.nature.com/doifinder/10.1038/nature01031>
- [4] JONES, Susan a Janet M. THORNTON. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*. 1996-01-09, vol. 93, no. 1, s. 13 – 20. ISSN 0027-8424. DOI: 10.1073/pnas.93.1.13. Dostupné z: <http://www.pnas.org/cgi/doi/10.1073/pnas.93.1.13>
- [5] CZJZEK, M et al. Crystal structure of a dimeric octaheme cytochrome c3 (Mr 26000) from *Desulfovibrio desulfuricans* Norway. *Structure*. 1996, roč. 4, č. 4, s. 395 – 404. ISSN 09692126. DOI: 10.1016/S0969-2126(96)00045-7. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S0969212696000457>
- [6] DOWNING, A. Kristina. *Protein NMR techniques*. 2nd ed. Totowa, NJ: Humana Press, 2004. ISBN 15-925-9809-9.
- [7] JANIN, Joël a Cyrus CHOTHIA. The structure of protein-protein recognition sites. *The Journal of Biological Chemistry*. 1990, vol. 265, no. 27, s. 16027 – 16030. ISSN 0021-9258.
- [8] MA, Buyong. Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences*. 2003, vol. 100, no. 10, s. 5772 – 5777. ISSN 0027-8424. DOI: 10.1073/pnas.1030237100. Dostupné z: <http://www.pnas.org/cgi/doi/10.1073/pnas.1030237100>
- [9] FERNÁNDEZ-RECIO, Juan. Prediction of protein binding sites and hot spots. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 2011, vol. 1, no. 5, s. 680 – 698. ISSN 17590876. DOI: 10.1002/wcms.45. Dostupné z: <http://doi.wiley.com/10.1002/wcms.45>
- [10] RAJAMANI, Deepa et al. Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*. 2004, vol. 101, no. 31, s. 11287 – 11292. ISSN 1091-6490. DOI: 10.1073/pnas.0401942101. Dostupné z: <http://www.pnas.org/cgi/doi/10.1073/pnas.0401942101>

- [11] CLACKSON, Tim a James A. WELLS. A hot spot of binding energy in a hormone-receptor interface. *Science*. 1995-01-20, vol. 267, no. 5196, s. 383 – 386. ISSN 0036-8075. DOI: 10.1126/science.7529940. Dostupné z: <http://www.sciencemag.org/cgi/doi/10.1126/science.7529940>
- [12] OZCABAN, Saliha E. A. Conformational ensembles, signal transduction and residue hot spots: Application to drug discovery. *Current opinion in drug discovery*. 2010, vol. 13, no. 5, s. 527 – 537. ISSN 2040-3437.
- [13] SHULMAN-PELEG, Alexandra et al. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biology*. 2007, vol. 5, no. 1, s. 43 – 53. ISSN 1741-7007. DOI: 10.1186/1741-7007-5-43. Dostupné z: <http://www.biomedcentral.com/1741-7007/5/43>
- [14] BOGAN, Andrew A a Kurt S. THORN. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*. 1998, vol. 280, no. 1, s. 1 – 9. ISSN 0022-2836. DOI: 10.1006/jmbi.1998.1843. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S0022283698918435>
- [15] OFRAN, Yanay a Burkhard ROST. Protein-Protein Interaction Hotspots Carved into Sequences. *PLoS Computational Biology*. 2007, vol. 3, no. 7, s. 1169 – 1176. ISSN 1553-734x. DOI: 10.1371/journal.pcbi.0030119. Dostupné z: <http://dx.plos.org/10.1371/journal.pcbi.0030119>
- [16] GROSDIDIER, Solène a Juan FERNÁNDEZ-RECIO. Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics*. 2008, vol. 9, no. 1, s. 447 – 459. ISSN 1471-2105. DOI: 10.1186/1471-2105-9-447. Dostupné z: <http://www.biomedcentral.com/1471-2105/9/447>
- [17] CHOTHIA, Cyrus a Joël JANIN. Principles of protein-protein recognition. *Nature*. 1975-8-28, vol. 256, no. 5520, s. 705 – 708. ISSN 0028-0836. DOI: 10.1038/256705a0. Dostupné z: <http://www.nature.com/doi/10.1038/256705a0>
- [18] *Alanine Scanning Energetics Database* [online]. 2001 [cit. 2013-05-09]. Dostupné z: <http://nic.ucsf.edu/asedb/>
- [19] BLUNDELL, Tom L. a S. PATEL. High-throughput X-ray crystallography for drug discovery. *Current Opinion in Pharmacology*. 2004, vol. 4, no. 5, s. 490 – 496. ISSN 1471-4892. DOI: 10.1016/j.coph.2004.04.007. Dostupné z: <http://linkinghub.elsevier.com/retrieve/pii/S1471489204001225>
- [20] BUDĚŠÍNSKÝ, Miloš a Jan PELNAŘ. *Fyzikálně-chemické metody: (nukleární magnetická rezonance) : 25. svazek cyklu Organická chemie*. Praha: Ústav organické chemie a biochemie Akademie věd ČR, 2000, 248 s. ISBN 80-862-4107-6.
- [21] WÜTHRICH, Kurt. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science*. 1989-01-06, vol. 243, no. 4887, s. 45

- 50. ISSN 0036-8075. DOI: 10.1126/science.2911719. Dostupné z: <http://www.sciencemag.org/cgi/doi/10.1126/science.2911719>
- [22] SHEN, Yang et al. Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences of the United States of America*. 2008-03-25, vol. 105, no. 12, s. 4685 – 4690. ISSN 1091-6490. DOI: 10.1073/pnas.0800256105. Dostupné z: <http://www.pnas.org/cgi/doi/10.1073/pnas.0800256105>
- [23] KORTEMME, Tanja a David BAKER. A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences*. 2002-10-29, vol. 99, no. 22, s. 14116 – 14121. ISSN 0027-8424. DOI: 10.1073/pnas.202485799. Dostupné z: <http://www.pnas.org/cgi/doi/10.1073/pnas.202485799>
- [24] TUNCBAG, Nurcan, Ozlem KESKIN a Attila GURSOY. HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Research*. 2010-06-24, vol. 38, s. 402 – 406. ISSN 0305-1048. DOI: 10.1093/nar/gkq323. Dostupné z: <http://www.nar.oxfordjournals.org/cgi/doi/10.1093/nar/gkq323>
- [25] SAHU, Sitanshu S. a Ganapati PANDA. Efficient Localization of Hot Spots in Proteins Using a Novel S-Transform Based Filtering Approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2011, vol. 8, no. 5, s. 1235 – 1246. ISSN 1545-5963. DOI: 10.1109/TCBB.2010.109. Dostupné z: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5611489>
- [26] KORTEMME, Tanja, David E. KIM a David BAKER. Computational Alanine Scanning of Protein-Protein Interfaces. *Science's STKE*. 2004-02-03, vol. 2004, issue 219. ISSN 1525-8882. DOI: 10.1126/stke.2192004pl2. Dostupné z: <http://stke.sciencemag.org/cgi/doi/10.1126/stke.2192004pl2>
- [27] CHENG, Tammy M.-K., Tom L. BLUNDELL a Juan FERNANDEZ-RECIO. PyDock: Electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*. 2007-08-01, vol. 68, no. 2, s. 503 – 515. ISSN 0887-3585. DOI: 10.1002/prot.21419. Dostupné z: <http://doi.wiley.com/10.1002/prot.21419>
- [28] YAN, Changhui et al. Characterization of Protein-Protein Interfaces. *The Protein Journal*. 2008, vol. 27, no. 1, s. 59 – 70. ISSN 1572-3887. DOI: 10.1007/s10930-007-9108-x. Dostupné z: <http://www.springerlink.com/index/10.1007/s10930-007-9108-x>
- [29] RAMACHANDRAN, Parameswaran a Andreas ANTONIOU. Localization of hot spots in proteins using digital filters. In: IEEE SIGNAL PROCESSING SOCIETY AND THE IEEE COMPUTER SOCIETY. *The 6th IEEE International Symposium on Signal Processing and Information Technology August 27-30, 2006, Listel Vancouver Hotel, Vancouver, BC, Canada*. Piscataway, N.J.: IEEE, 2006, s. 926 – 931. ISSN 0-7803-9754-1.

- [30] VELJKOVIĆ, V. et al. Is it Possible to Analyze DNA and Protein Sequences by the Methods of Digital Signal Processing?. *IEEE Transactions on Biomedical Engineering*. 1985, vol. BME-32, no. 5, s. 337 – 341. ISSN 0018-9294. DOI: 10.1109/TBME.1985.325549. Dostupné z: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4122061>
- [31] RAMACHANDRAN, Parameswaran a Andreas ANTONIOU. Identification of Hot-Spot Locations in Proteins Using Digital Filters. *IEEE Journal of Selected Topics in Signal Processing*. 2008, vol. 2, no. 3, s. 378 – 389. ISSN 1932-4553. DOI: 10.1109/JSTSP.2008.923850. Dostupné z: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4550565>
- [32] STOCKWELL, R. G. et al. Localization of the complex spectrum: the S transform. *IEEE Transactions on Signal Processing*. 1996, vol. 44, no. 4, s. 998 – 1001. ISSN 1053587x. DOI: 10.1109/78.492555. Dostupné z: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=492555>
- [33] THE UNIPROT CONSORTIUM. *UniProt* [online]. 2002, 2013-05-01 [cit. 2013-05-09]. Dostupné z: www.uniprot.org
- [34] SIMON, Carine et al. The S-Transform and Its Inverses: Side Effects of Discretizing and Filtering. *IEEE Transactions on Signal Processing*. roč. 55, č. 10, s. 4928-4937. ISSN 1053-587x. DOI: 10.1109/TSP.2007.897893. Dostupné z: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4305459>
- [35] TUNCBAG, N., A. GURSOY a O. KESKIN. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*. 2009-05-28, vol. 25, issue 12, s. 1513-1520. DOI: 10.1093/bioinformatics/btp240. Dostupné z: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp240>

Seznam zkratek

ASEdb	Alanine Scanning Energetics Database
ASM	alanine scanning mutagenesis – mutageneze alaninovým vyhlazením
AUC	area under curve – plocha pod křivkou
CWT	continuous wavelet transform – spojitá vlnková transformace
EIIP	electron-ion interaction potencial
IIR	infinite impulse response
ISIS	interaction sites identified from semence
IST	inverzní S-transformace
MAPPIS	multiple alignment of protein-protein interfaces
mRNA	mediátorová RNA
NIP	normalized interface propensity – normalizovaná četnost rezidua
NMR	nukleární magnetická rezonance
PDB	Protein Data Bank
PEII	potencial of electron-ion interaction
PPV	positive predictive value – pozitivní prediktivní hodnota
ROC	receiver operating characterictic
RRM	resonant recognition model – model rezonančního rozpoznání
SASA	solvent accessible surface area – povrch přístupný rozpouštědлу
SE	senzitivita
snRNA	small nuclear RNA – malá jaderná RNA
SP	specificita
ST	S-transformace
STFT	short time Fourier transform – krátkodobá Fourierova transformace