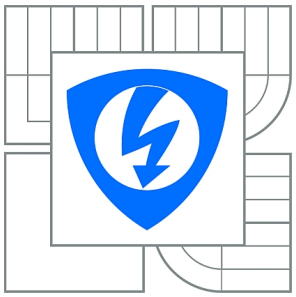
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

METODY PREDIKCE SEKUNDÁRNÍ STRUKTURY PROTEINŮ

METHODS FOR PREDICTION OF SECONDARY STRUCTURE IN PROTEINS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

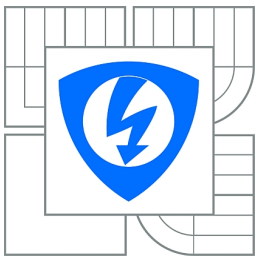
AUTOR PRÁCE — NINA HOŠTÁKOVÁ

AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ

SUPERVISOR

BRNO 2012



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Studentka: Nina Hoštáková ***ID:*** 120113
Ročník: 3 ***Akademický rok:*** 2011/2012

NÁZEV TÉMATU:

Metody predikce sekundární struktury proteinů

POKYNY PRO VYPRACOVÁNÍ:

1) Nastudujte problematiku 1D, 2D a 3D struktur, které vytváří sekvence aminokyselin. Vypracujte literární rešerši metod používaných pro predikci a zkoumání prostorových struktur řetězců aminokyselin. 2) Srovnejte vlastnosti, výhody a nevýhody jednotlivých metod. 3) Nalezněte na internetu volně přístupný software pro predikci sekundárních proteinových struktur, software vyzkoušejte na vybraném souboru dat. 4) V programovém prostředí Matlab realizujte vybranou metodu predikce sekundární struktury proteinů. 5) Realizovanou metodu vyzkoušejte na souboru dat a výsledky porovnejte s výsledky volně dostupného softwaru.

DOPORUČENÁ LITERATURA:

[1] FASMAN, G.D. Prediction of Protein Structure and the Principles of Protein Conformation. Springer, 1989.
[2] CHOU, P.Y., FASMAN, G.D. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. Relat. Areas Mol. Biol. Vol. 47, pp. 45-148, 1978.

Termín zadání: 6.2.2012 ***Termín odevzdání:*** 25.5.2012

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti bakalářské práce:

prof. Ing. Ivo Provazník, Ph.D.
Předseda oborové rady

UPOZORNĚNÍ:

Abstrakt

Zkoumání proteinových struktur má klíčový význam při zjišťování způsobů působení bílkovin v organizmu. Práce zpracovává problematiku 1D, 2D a 3D struktur, do kterých se proteiny v prostoru uspořádavají. Důraz je kladen na sekundární strukturu, kterou lze predikovat přímo ze sekvencí aminokyselin a následně ji využít pro odhad prostorové struktury. Tomuto postupu se věnují výpočetní metody, které pomocí algoritmů konvertují sled aminokyselin na sled sekundárních struktur. Přímým určením struktury, pomocí vytváření strukturních modelů, se zabývá část věnovaná experimentálním metodám (NMR spektroskopiie, RTG krystalografie). Hlavním cílem práce je programová realizace metody predikující sekundární strukturu proteinů. Vytvořený program je doplněn o grafické uživatelské rozhraní. Výsledky programu, navrženého na základě metody Chou-Fasman, jsou v závěrečné části práce porovnány s výstupy volně dostupných softwarů z internetu.

Klíčová slova

proteiny, sekundární struktura, predikce, RTG krystalografie, NMR spektroskopie, výpočetní metody, Chou - Fasmanova metoda

Abstract

The examination of protein structure is crucial in determining protein function in organism. This work deals with the issue of 1D, 2D and 3D structures, into which are proteins organized in space. Emphasis is placed on secondary structure, which can be predicted directly from the amino acid sequences and then used for the estimation of spatial structure. On this procedure are focused computational methods, using algorithms that convert the order of amino acids into the order of preferences for secondary structures. To direct determination of the structure by creating structural models is devoted chapter Experimental Methods (NMR spectroscopy, RTG crystallography). The main aim of this work is practical realization of protein secondary structure prediction method. The created program is supplemented by graphical user interface. In the final part the results of the program based on Chou-Fasman method are compared to the outputs of freely available softwares from the Internet.

Keywords

proteins, secondary structure, prediction, RTG crystallography, NMR spectroscopy, computational methods, Chou-Fasman method

Bibliografická citace

HOŠTÁKOVÁ, N. *Metody predikce sekundární struktury proteinů:* bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2012. 54 s., 1 příl. Vedoucí práce Ing. Denisa Maděránková.

Prohlášení

Prohlašuji, že svou bakalářskou práci na téma *Metody predikce sekundární struktury proteinů* jsem vypracovala samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla 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 § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 25.5.2012 .....
podpis autorky

Poděkování

Rada by som sa poďakovala vedúcej práce Ing. Denise Maděránkovej za účinnú metodickú pomoc, odborné vedenie a cenné rady k spracovaniu práce na konzultáciách.

V Brně dne 25.5 .....
podpis autorky

Obsah

Úvod

Proteíny predstavujú základ živého organizmu a súvisia takmer s každým prejavom jeho existencie. Pre pochopenie priebehu zložitých biologických a chemických procesov v organizmoch je nutné určiť, akým spôsobom v ňom jednotlivé proteíny participujú. Rozhodujúci význam pri zisťovaní funkcie a spôsobu účinku bielkovín v organizme má práve určovanie ich štruktúry. Štruktúra proteínu je kľúčom k odhaleniu vzťahu medzi génom, vyjadreným postupnosťou aminokyselín, a jeho prejavom – funkciou proteínu.

Štruktúru je možné „rekonštruovať“ z izolovaných vzoriek proteínov pomocou experimentálnych metód. Získané modely proteínov poskytujú veľmi presnú štruktúrnu informáciu, vyžadujú však sofistikované vybavenie a sú finančne a časovo náročné. Rôzne obmedzenia, vyplývajúce z charakteru používaných metód, tiež spôsobujú, že nie sú aplikovateľné na všetky proteíny. Ich využitie nie je preto pre požiadavky modernej biológie úplne postačujúce.

Rýchlo sa rozvíjajúca genomika a metódy sekvenovania DNA prispeli k odlišnému prístupu v určovaní bielkovinovej štruktúry. Nastal stav, kedy množstvo známych proteínových sekvencií vysoko prevyšuje rýchlosť určovania ich piestorovej štruktúry. Z tohto dôvodu je v súčasnej dobe prevažujúca snaha derivovať štruktúrnu informáciu priamo z aminokyselinových sekvencií, tj. predikovať štruktúru. Touto úlohou sa zaoberajú výpočtové predikčné metódy, ktoré predstavujú hlavnú náplň tejto práce.

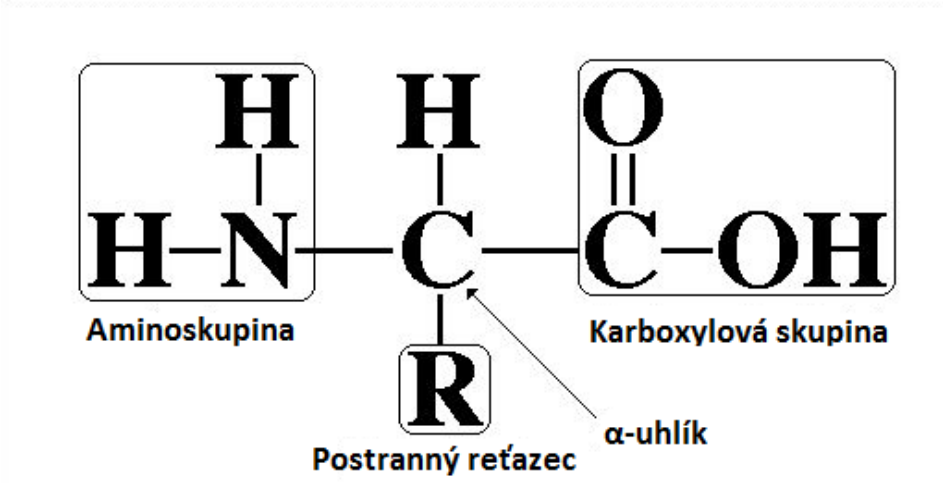
Práca sa zameriava špeciálne na predikciu sekundárnej štruktúry, ktorá je dôležitým krokom pri odhade štruktúry na vyšších úrovniach usporiadania (3D, 4D). Cieľom je programová realizácia metódy 2D predikcie a jej otestovanie na vybranom súbore dát. Prvá časť dokumentu je venovaná všeobecnému popisu proteínovej problematiky – funkcii, zloženiu a usporiadaniu proteínov v priestore. V ďalšej časti sa práca zaoberá samotnými metódami skúmania týchto štruktúr – experimentálnymi s využitím RTG, NMR a výpočtovými, zameranými na sekundárnu štruktúru. Nasleduje popis realizácie predikčného programu vrátane grafického rozhrania pre užívateľa. Na záver je uvedený prehľad vybraných softvérov dostupných z internetu, ktoré sa zaoberajú touto problematikou, zobrazenie výsledkov ich predikcií a vzájomné porovnanie s realizovaným programom.

1 Proteíny

Proteíny, z chemického hľadiska organizovanejšia forma polypeptidov, predstavujú biopolyméry aminových kyselín so širokospektrálnym významom v živých štruktúrach. Zabezpečujú uchovávanie a prenos rôznych častíc, vrátane makromolekúl na strane jednej a elementárnych častíc na strane druhej. Vo forme enzýmov katalyzujú biologické metabolické procesy. Zúčasťňujú sa fotosyntézy, kde bielkovinové komplexy sprostredkúvajú priebeh chemických reakcií a vedú tok elektrónov. Ďalšie druhy sa podieľajú na štruktúre filament a bunkových membrán. Tu zároveň pôsobia ako prenášače molekúl z extracelulárneho do intracelulárneho prostredia. Na vyššej úrovni sú súčasťou kostí, vlasov a svalov – umožňujú premenu chemickej energie na energiu mechanickú. Predstavujú základný zdroj dusika v organizme, ďalej využívaného pre syntézu proteínov a nukleových kyselín *de novo*. V neposlednom rade zastávajú funkciu informačnú (hormóny), imunitnú (rozpoznávania patogénov) a signálnu. [1], [2]

1.1 Chemické zloženie proteínov

Aj keď majú proteíny veľkú funkčnú variabilitu, z molekulárneho hľadiska sú si veľmi podobné. Ich rozdiely vo funkcii sú spôsobené predovšekým rôznymi usporiadaniami a štruktúrou v priestore. Chemické zloženie proteínov je teda prakticky rovnaké, sú tvorené aminokyselinovým reziduami. Jedná sa o kombinácie 20[1] možných aminokyselín - monomérov, ktoré sa navzájom viažu peptidovou väzbou a v zoskupení spravidla 50 až 3000 vytvárajú polypeptidové reťazce. Takéto reťazce sú lineárne a opakuje sa v nich základná jednotka spoločná pre všetky aminokyseliny. V priemere je polymérny reťazec tvorený asi 250-timi aminokyselinovými reziduami. Molekulová hmotnosť je teda značne vysoká, od 10 000 po 50 000 kDa a dĺžka reťazca dosahuje až 3 μm. [3], [5]



Obr. 1: Obecný chemický vzorec aminokyselín, centrálny atóm uhlíka je označovaný ako α-uhlík [4]

1 V súčasnej dobe sa počet geneticky kódovaných aminokyselín rozšíril na 21. Poslednú zatiaľ známu aminokyselinu podieľajúcu sa na štruktúre proteínov predstavuje selenocysteín.

1.1.1 Aminokyseliny

Aminokyseliny, základný stavebný kameň bielkovín, sa zaraďujú medzi organické karboxylové kyseliny. Okrem charakteristickej karboxylovej skupiny –COOH však navyše obsahujú primárnu aminoskupinu $–NH_2$. Táto skupina je viazaná prevažne na α-uhlík priamo spojený s karboxylovou skupinou (obr. 1). Na rovnaký uhlík je viazaný i postranný reťazec, ktorým sa jednotlivé aminokyseliny líšia. Rôzne postranné reťazce spôsobujú odlišnosti vo fyzikálno- chemických vlastnostiach. Väčšinu aminokyselín podieľajúcich sa na tvorbe proteínov označujeme ako α- aminokyseliny. Výnimkou je prolín, ktorý má postranný reťazec zviazaný s dusíkom charakteristickej skupiny. Tá je tým pádom sekundárna a prolín vytvára α - iminoskupinu [2], [3]. Aminokyseliny sa vyznačujú asymetrickým, tzv. chirálnym uhlíkom, ktorý stáča rovinu lineárne polarizovaného svetla. Jedná sa o spomínaný α-uhlík, ktorý viaže štyri rôzne substituenty. Výnimkou je glycín, pretože jeho postranný reťazec tvorí samotný atóm vodíka, ktorý sa nepovažuje za substituent. Chirálny uhlík spôsobuje, že aminokyseliny sa vyskytujú vo forme L- izomérov či D- izomérov. Pre proteíny je charakteristická L- izoméria. Toto označenie je odvodené od usporiadania, v akom sa vyskytuje aminokyselina vo Fisherovej projekcii [3]. Pre lepšiu prehľadnosť sa aminokyseliny označujú kódom, ktorý predstavuje trojpísmenovú, prípadne jednopísmenovú skratku ich triviálneho názvu (Tab. 1). [5]

Tab. 1: Prehľad aminokyselín a ich skrátené označenia [6]

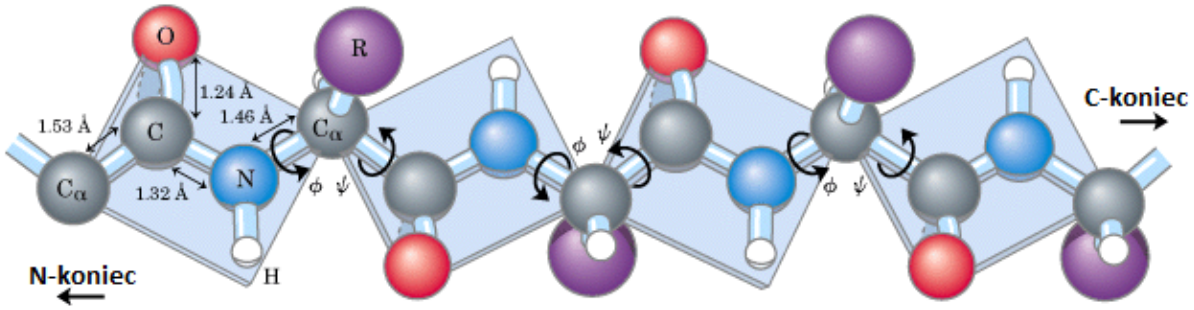
Značenie aminokyselín					
Triviálny názov	3-písm.	1-písm.	Triviálny názov	3-písm	1-písm.
Alanín	Ala	A	Glycín	Gly	G
Valín	Val	V	Serín	Ser	S
Leucín	Leu	L	Treonín	Thr	T
Izoleucín	Ile	I	Cysteín	Cys	C
Fenylalanín	Phe	F	Tyrozín	Tyr	Y
Tryptofan	Trp	W	Asparagín	Asn	N
Metionín	Met	M	Glutamín	Gln	Q
Prolín	Pro	P	Lyzín	Lys	K
Asparová kys.	Asp	D	Arginín	Arg	R
Glutamová kys.	Glu	E	Histidín	His	H

Hydrofóbne Hydrofilné Kyslé Zásadité

1.1.2 Peptidové väzby

V proteínoch sa dve a viac aminokyselín spája do dlhších reťazcov pomocou kovalentných peptidových väzieb. Väzba vzniká procesom kondenzácie, kde sa aminoskupina jednej aminokyseliny spája s karboxylovou skupinou druhej aminokyseliny a vylučuje sa molekula vody. Vzniknutá

peptidová väzba má potom tvar –CO–NH–. Značenie vzniknutých peptidov sa následne odvodzuje od jedného z dvoch koncov; N-koniec vytvára aminokyselina s nezreagovanou aminoskupinou, C- koniec aminokyselina s nezreagovanou karboxylovou skupinou. Konvencia zvyčajne udáva poradie aminokyselín od N-konca [3]. Peptidové väzby sú veľmi stabilné vo vodnom prostredí s neutrálnym pH a prípadná hydrolýza je enzymaticky kontrolovaná. Dôležitým faktorom v stabilite a ďalšom správaní polypeptidu je rezonancia, tj. kmitanie elektrónov ponad atómy väzby. Tento jav spôsobuje jednak zvýšenú polaritu väzby (prejavuje sa dipólovým momentom) a jednak čiastočný dvojitý charakter peptidovej väzby (40%). Tento fakt má za následok rovinné usporiadanie troch atómov – karbonylového uhlika C, karbonylového kyslíka O a amidového dusíka N. Tým je obmedzená voľná rotácia atómov okolo väzby a teda aj možnosti priestorového usporiadania. Zvyšné dve väzby v kostre peptidu (medzi atómami centrálneho uhlíka C_α, dusíka a karbonylového uhlíka) sú väzby jednoduché s možnou rotáciou. Celkové usporiadanie väzby v priestore vyjadrujeme pomocou uhlov rotácie, tzv. torzných uhlov (Φ, Ψ, Ω). Torzný Φ uhol je definovaný ako uhol väzby N-C_α k priľahlej peptidovej väzbe a Ψ uhol medzi väzbou C-C_α a priľahlou peptidovou väzbou (Obr. 2). Uhol Ω, ktorý má poväčšine hodnotu 180°, spôsobuje spomínanú planaritu peptidovej väzby. V takomto prípade sa jedná o konfiguráciu *trans*. Pomerne ojedinená konfigurácia peptidovej väzby *cis* sa vyskytuje pri nulovou uhle Ω, známa napríklad u rezidua prolínu. Komplexne si môžeme predstaviť kostru polypeptidového reťazca ako striedajúce sa kovalentné väzby, umožňujúce rotáciu, s pevnými planárnymi peptidovými väzbami. Takýto charakter hlavného reťazca značne obmedzuje počet možných usporiadaní proteínových makromolekúl v priestore. [4], [6]



Obr. 2: Viazanie aminokyselín peptidovými väzbami do peptidového reťazca s vyznačenými torznými uhlami [22]

2 Štruktúry proteínov

2.1 Primárna štruktúra

Vlastnosti jednotlivých proteínov sú dané nielen zastúpením aminokyselín, ale tiež ich usporiadaním v reťazci. Primárna štruktúra teda udáva počet a sled aminokyselinových reziduí v bielkovine. Množstvo rôznych proteínov, ktoré môžu vznikať z rovnakých molekúl aminokyselín je značné. Matematicky je tento počet vyjadrený permutáciou počtu aminokyselín v proteíne. Ako dva odlišné prípady berieme aj výskyt aminokyseliny na C a N konci. [5]

Primárna štruktúra je teda daná hlavným polypeptidovým reťazcom a postrannými reťazcami aminokyselín. V hlavnom reťazci proteínov sa vyskytujú tri základné atómy každej aminokyseliny, ktoré sa pravidelne opakujú. Jedná sa o amidový dusík (N), centrálny atóm uhlíku (C_α) a karbonylový uhlík (C). [3]

Postranné reťazce môžeme zaradadiť do nasledujúcich skupín (Tab. 1):

- nepolárne – tieto reziduá sú hydrofóbne, vyhýbajú sa interakciám s vodou. Spájaju sa iba slabými van der Waalsovými väzbami, príkladom sú aminokyseliny obsahujúce skupinu $–CH_3$
- polárne – označované ako hydrofilné, môžu obsahovať neutrálne alebo nabité skupiny (– COOH, – OH). Hydrofília sa prejavuje v tendencii vytvárať interakcie s vodou, prípadne inými hydrofilnými molekulami, väčšinou vo forme vodíkových mostíkov. Skupiny s nábojom ďalej rozdeľujeme na:
    - acidické (záporne nabité) – obsahujú početné dikarboxylové skupiny
    - bázické (kladne nabité) – obsahujú početné diamínové skupiny
- amfipatické – tieto reťazce môžu nadobúdať za rôznych podmienok rôzne vlastnosti, prípadne majú čiastočne polárny a čiastočne nepolárny charakter (Trp, Tyr, Lys, Met). [6], [7]

Primárna štruktúra určuje biologické a chemické vlastnosti proteínu a následne aj formu vyššieho usporiadania. Je tomu tak preto, že proteíny sa z rozvinutého primárneho reťazca usporiadajú do svojej natívnej formy samovoľne. Na základe primárnej štruktúry je tiež možné zisťovať vzájomné vzťahy s inými proteínmi v evolúcii. 1D štruktúra je vyjadrením genetickej informácie uloženej v DNA či RNA spracovanej procesom transkripcie a translácie. Poradie aminokyselín v polypeptidovom reťazci zodpovedá poradiu nukleových kyselín daných genetickým kódom. Primárna štruktúra podáva kompletnú informáciu o konfigurácii proteínu, ale naopak, bez ďalšieho skúmania sa z nej nedá dozvedieť nič o jeho konformácii. [3]

2.2 Sekundárna štruktúra

Sekundárna štruktúra proteínov predstavuje lokálne usporiadanie polypeptidového reťazca. Obsahuje pravidelne sa opakujúce charakteristické segmenty, ktorých tvar a štruktúru udávajú Φ a Ψ torzné uhly.

Proteín je schopný vytvárať 2D štruktúrne elementy vďaka vodíkovým mostíkom, ktoré ich stabilizujú. Zároveň neutralizujú polárny charakter hlavného reťazca a ten tak dokáže existovať v hydrofóbnom prostedí jadra štruktúry. Zo samotnej sekundárnej štruktúry je možné v určitých prípadoch dokonca priamo determinovať funkciu (α-hélixy v DNA). [1], [6]

2.2.1 Helikálne štruktúry

Základným druhom sekundárnej štruktúry sú tzv. hélixy alebo závitnice. Majú osobitné názvoslovie odvodené od charakteristických geometrických parametrov. Medzi najzákladnejšie patria: počet aminokyselinových reziduí na jeden závit, počet atómov v sľučke uzatvorenej susednými vodíkovými väzbami, výška závitu a stúpanie závitnice. Prvé dve charakteristiky udávajú označenie helixu. Najznámejšia α-závitnica má napríklad označenie 3.6_{13}. Hélixy sa vytvárajú stočením polypeptidovej kostry do tvaru závitu okolo uhlíka C_α v konštantnej vzdialenosti. Vzniknutá závitnica je chirálna, túto chiralitu však nemožno stotožnovať s chiralitou aminokyselín. Rozoznávame závitnice ľavotočivé a pravotočivé. α- aminokyseliny, z ktorých proteíny pozostávajú, však vo väčšine prípadov nie sú schopné vytvárať ľavotočivé závitnice. Je to najmä kvôli blízkosti väzieb postranných reťazcov a karbonylovej skupiny. Konformácia väčšiny hélixov je taká, že hydrofilné skupiny aminokyselín sa vyskytujú na povrchu, kým hydrofóbne skupiny sú vnorené do vnútra závitu.

- **α-hélix**

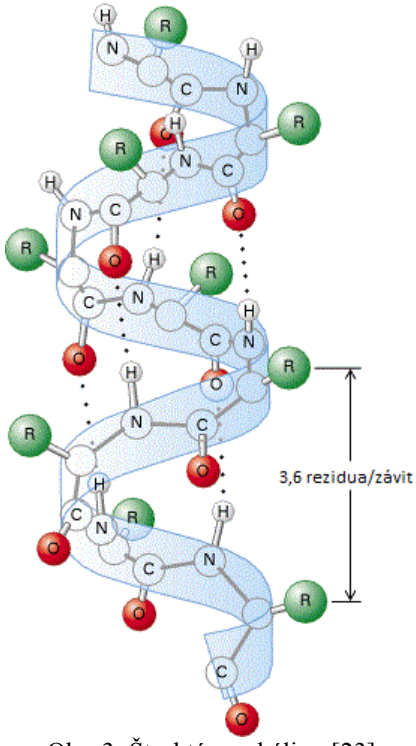
Predstavuje najrozšírenejšiu helikálnu štruktúru. Z konformačného hľadiska sa jedná o pravotočivú závitnicu s presne definovanými torznými uhlami; Φ = 57° a Ψ = - 47°. α-hélixy sú tvorené 3,6 reziduami na závit. Vytvárajú vodíkové mostíky medzi karbonylom (C=O) *n*- tého rezidua a amidovou skupinou (N – H) *n+4* -tého rezidua. Vzorec viazania molekúl sa preto označuje *n+4*. Mostíky sú usporiadané rovnobežne s osou hélixu a teda v súlade s dipólovým momentom peptidovej kostry. Preto sú α-hélixy najstabilnejšie, energeticky najvýhodnejšie a vyskytujú sa ako v globulárnych, tak vo fibrilárnych proteínoch. Spomínaným spôsobom sú spojené všetky reziduá s výnimkou prvého amidu a posledného karbonylu závitnice (Obr. 3).

- **Špeciálne helikálne štruktúry**

Medzi najčastejšie patria modifikácie klasickej závitnice, ktoré označujeme ako štruktúry *n+3* (hélix 3_{10}) a *n+5* (hélix π), podľa vzorca spájania reziduí vodíkovými väzbami.

Hélix 3_{10} je od α-hélixu tvarovo značne odlišný, je užší a pretiahnutý, jeden závit má tri peptidové reziduá. Reťazec je vďaka tomu pevnejšie utiahnutý. Táto štruktúra často vzniká medzi α- hélixom a zvyšnou časťou polypeptidového reťazca, alebo tvorí ukončenie reťazca na C-konci.

Hélix π má, naopak, širšiu a málo zvinutú konformáciu s 4,4 jednotkami na závit. Táto štruktúra je veľmi málo stabilná. Kvôli sploštenému usporiadaniu sa vytvára v smere osi polypeptidu otvor, ktorý spôsobuje, že nemôže dochádzať k stabilizácii van der Waalsovými väzbami. Preto sú tieto štruktúry v proteínoch veľmi vzácne. [7]

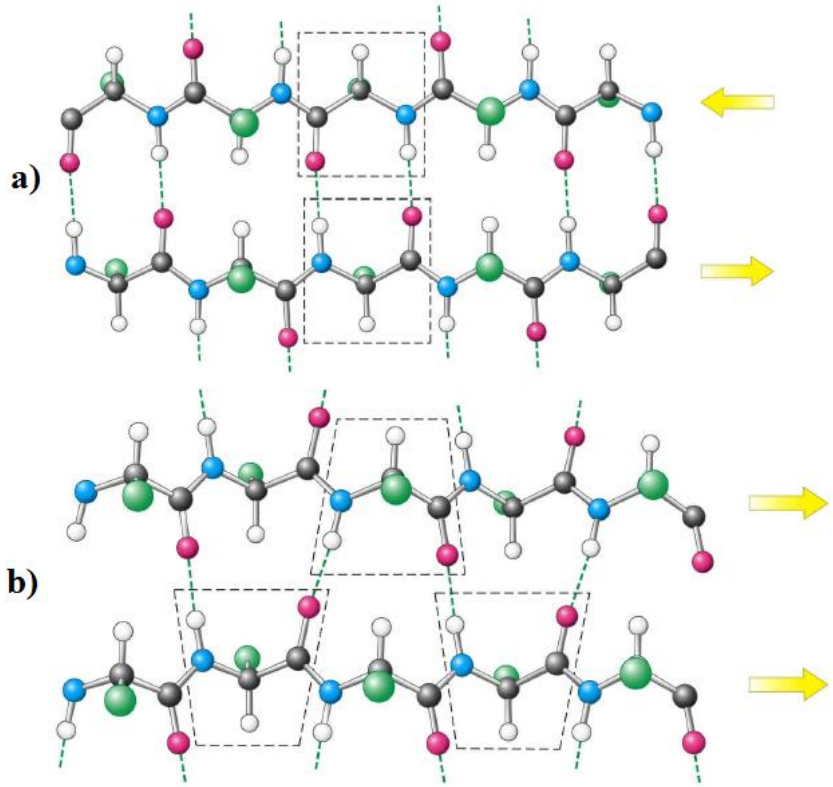


Obr. 3: Štruktúra α-hélixu [23]

2.2.2 β-štruktúry

Narozdiel od hélixov, vodíkové väzby nespájajú reziduá vo vnútri ß-štruktúr, uplatňujú sa len pri spájaní priľahlých reťazcov. β-štruktúry rozdeľujeme na paralélne a antiparalélne. Často sú označované ako ß-skladané listy podľa tvaru konformácie, ktorý zaujímajú. Paralélne listy predstavujú vlákna spojené vodíkovou väzbou v rovnakom smere, tj. C-koniec jedného vlákna je spojený s N- začiatkom ďalšieho vlákna. V prípade antiparalélnych listov sa vlákna spájajú v opačnom smere . Na prvé vlákno v klasickom smere N-C sa pripája ďalšie vlákno v opačnom smere, C- koncom (Obr. 4). Oba druhy majú charakteristický spôsob spájania vodíkovými väzbami. Antiparalélne listy sa navzájom spájajú väzbou v rovine listov (β-ohyb), spojenie paralélnych vlákien je o niečo zložitejšie. Spoj sa vytvára priečnou väzbou, ktorá už zviera s rovinou listu určitý uhol. Paralélne štruktúry sú kvôli tomu menej stabilné v priestore. Oba β-štruktúrne elementy sa môžu rôznym spôsobom kombinovať a vytvárať tak zmiešané štruktúry. Postranné reťazce β-listov vybiehajú striedavo od stredu osi reťazca nad a pod rovinu listu. Listy nemajú úplne rozvinuté usporiadanie, aj keď by sa tak na prvý pohľad mohlo zdať. Torzné uhly Φ a Ψ sa pohybujú okolo hodnôt -130° a +125°. To je dôvodom, prečo sú kostry reťazcov týchto štruktúr mierne pravotočivé, predovšetkým

u antiparalélnych listov. Koncové časti C a N sa už neviažu v rámci štruktúry, ale vytvárajú vodíkové mostíky buď s okolitým prostredím rozpúšťadla, napajajú sa na helikálne štruktúry alebo vytvárajú rozšírené β-štruktúry. [6], [7]



Obr. 4: β skladané listy: a) antiparalélne usporiadanie b) paralélne usporiadanie [19]

2.2.3 Nerepetitívne štruktúry

Zaraďujeme ich medzi doplňujúce štruktúry polypeptidu, ktoré majú tvar rôznych ohybov a sľučiek. Väčšinou sa podieľajú na spojoch medzi jednotivými α- a β-štruktúrami alebo ich ukončeniach. Majú variabilné geometrické usporiadanie a vyskytujú sa prevažne na povrchu proteínovej molekuly, pričom α- a β- štruktúry tvoria ich jadro.

Najznámejším príkladom je β-ohyb, označovaný tiež β-otočka či spätný ohyb. Vodíková väzba spája karbonyl *n*-tého rezidua s amidovou skupinou *n+3*-tého rezidua v spätnom smere. K zostaveniu tejto štruktúry teda stačia štyri (v určitých prípadoch iba tri) reziduá. Vďaka tomu predstavuje najjednoduchšiu sekundárnu štruktúru. Prípady, že štruktúra zasahuje len tri reziduá, sú ojedinelé, pretože takáto konformácia je už príliš „napnutá“. Väčšina ukončujúcich ohybov je vystavená vodnému prostrediu kvôli jednoduchšiemu vytvoreniu vodíkových väzieb. Vďaka

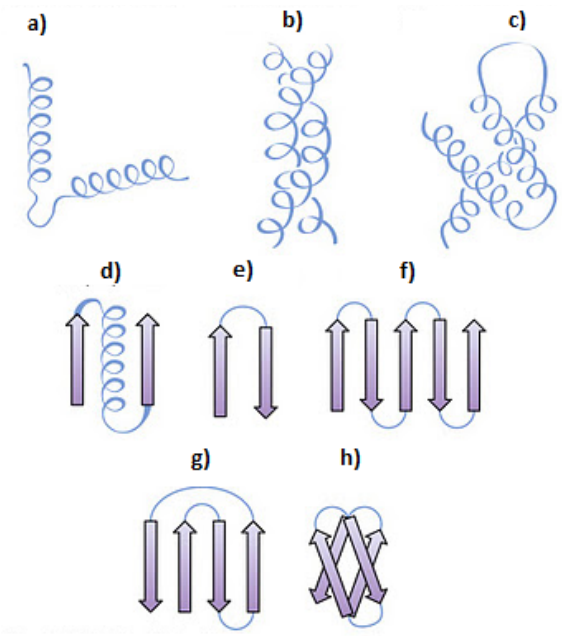
reverznému smeru reťazca umožňuje β-ohyb kompaktné usporiadanie proteínovej makromolekuly. [6], [7]

Ďalšiu známu nerepetitívnu štruktúru predstavuje Ω-sľučka, nazvaná podľa svojho zvinutého tvaru. Predpokladá sa, že zastáva úlohu v signálnych procesoch a pri špecifickom rozpoznávaní molekúl, pretože sa takmer výhradne nachádza na povrchu proteínov. [7]

2.3 Supersekundárna štruktúra

Štruktúrne motívy, ktoré tvoria charakteristické frekventované zoskupenia jednotlivých sekundárnych štruktúr, môžeme považovať za ďalšiu úroveň proteínovej štruktúry. Často bývajú asociované s konkrétnou funkciou proteínu.

Najjednoduchšou supersekundárnou štruktúrou je hélix- ohyb- hélix, známy z viazania kalcia či DNA. Obdobu v β-forme predstavuje β-vlásenka („β-hairpin“). Štruktúry β-α-β sú tvorené dvoma paralélnymi β-skladanými listami, ktorých spojenie je pomerne náročné. Často je realizované práve pomocou α-štruktúry a dvoch slučiek, ktoré spájajú jednotlivé elementy.



Obr. 5: Vybrané štruktúrne motívy: a) Helix-ohyb-helix b) Superhelix c) Helixový zväzok („Helix bundle“) d) β-α-β e) β - vlásenka f) β - meander g) Grécky kľúč h) β-sandwich [24]

Motív Grécky kľúč, naopak, pozostáva zo štyroch antiparalélnych β-vlákien. Priľahlé dve a dve vlákna sa spájajú β-ohybmi a vytvárajú tzv. β-meander. Zároveň sú navzájom spojené ďalším β - ohybom (Obr. 5).

Superhélix („coiled coil“) tvorí základ fibrilárnych proteínov. Vzniká vzájomným zvinutím dvoch α- hélixov okolo seba do jedného spoločného hélixu a umožňuje maximálne interakcie postranných reťazcov. Dochádza tým k výraznej stabilizácii štruktúry oproti samotným izolovaným α- hélixom. Vytvorením tejto štruktúry sa mení repetitívny vzorec aminosekvencií. Väzby vznikajú pravideľne po siedmich reziduách (heptátové opakovanie).

β-súdok („β-barrel“) je tvorený ôsmimi antiparalélnymi β-listami, ktoré sú navzájom pospájané ohybmi. Začiatočný a koncový list sa spája vodíkovou väzbou, čo spôsobuje charakteristickú guľovitú štruktúru. Má klasické usporiadanie postranných reťazcov, kde hydrofóbne reziduá vypĺňajú vnútro štruktúry a hydrofilné reziduá sú exponované na povrchu.

β-sendvič je tvorený dvoma listami oproti sebe, narozdiel od „barrelu“ konce nie sú spojené a vytvárajú tak štruktúru podobnú sendviču. Listy sú zväčša oproti sebe stočené o určitý uhol (približne 30°). Jednotlivé vlákna β-listov môžu mať rôznu topológiu, často sa vyskytuje Grécky kľúč.

Väčšina motívov je usporiadaná tak, že priľahlé elementy sú tvorené z aminokyselín, ktoré v primárnej štruktúre nasledujú za sebou. [1], [3], [6], [8]

2.4 Terciárna štruktúra

Terciárna štruktúra predstavuje celkové usporiadanie jedného polypeptidového reťazca v priestore. Tvoria ju tzv. domény, ktoré vznikajú buď asociáciou viacerých štrukturálnych motívov alebo priestorovým zvinutím samostatného polypeptidového reťazca. Domény sú základnými priestorovými a funkčnými jednotkami proteínu. Motívy, ktoré sú tvorené z priľahlých aminokyselín v primárnej štruktúre sa usporadúvajú do tercíárnej štruktúry, v ktorej spolu taktiež susedia.

Zjednodušene si môzno terciárnu štruktúru predstaviť ako sekvenciu jednotlivých štruktúrnych motívov. Ďalšia klasifikácia terciárnej štruktúry sa odvíja od priestorového usporiadania. Domény 3D štruktúry sú zaraďované do troch tried: α-domény, β-domény a α/β domény. α- a β- domény sú tvorené výhradne elementami svojho druhu, kým α/β domény obsahujú paralélny β- list obklopený α- hélixami. Za ďalšiu skupinu je možné považovať α+β domény, kde sa jednotlivé motívy vyskytujú oddelene, bez vzájomných asociácií. [8]

2.5 Kvartérna štruktúra

Jednotlivé polypeptidové reťazce zvyknú ďalej vzájomne interagovať. Vzniknuté komplexy reprezentujú kvartérnu štruktúru proteínu. Podjednotky štruktúry sú označované monoméry a proteín, pozostávajúci z viac ako jedného polypeptidového reťazca, oligomér, resp. polymér. [6]

3 Metódy skúmania proteínových štruktúr

3.1 Experimentálne metódy

Experimentálne metódy využívajú pre odhad priestorovej štruktúry relatívnu pozíciu jednotlivých atómov v molekule. K najrozšírenejším biofyzikálnym metódam skúmania bielkovinovej štruktúry patrí RTG kryštalografia a NMR spektroskopia. Fyzikálne sú tieto metódy odlišné, RTG kryštalografia zobrazuje difrakčný obrazec vzniknutý rozptylom žiarenia na elektrónoch, preto je obmedzená len na určenie nevodíkových atómov. Naopak, NMR zisťuje vzájomné pôsobenie jadier, je schopná určiť polohu všetkých atómov molekuly. Obe metódy teda podávajú na túto problematiku do istej miery komplementárny pohľad, vo výsledku ale predstavujú veľmi presnú formu stanovenia štruktúry. [6], [9]

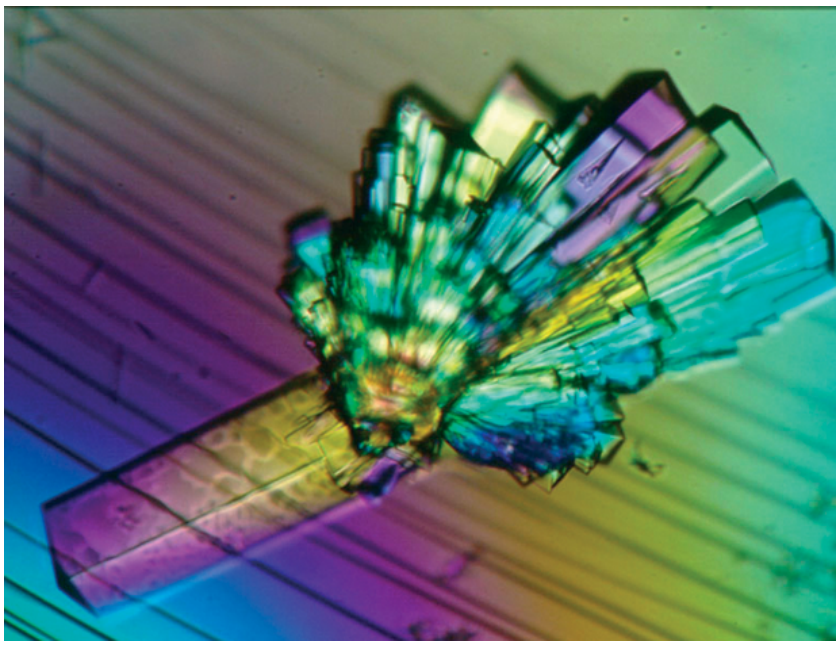
3.1.1 Rentgenová kryštalografia

Vnútorná štruktúra ideálneho kryštálu je daná určitým opakujúcim sa trojrozmerným atómovým motívom. Tento vzor je možné určiť na základe rozptylu sekundárneho žiarenia na elektrónoch kryštálu ožiarenom monochromatickým zväzkom X-žiarenia. Rentgenové žiarenie sa využíva preto, lebo má priaznivú vlnovú dĺžku, ktorá je porovnateľná s dĺžkou kovalentných väzieb. [19]

Prvým, časovo najnáročnejším, krokom k určeniu štruktúry je kryštalizácia makromolekuly, predpokladom dobrého výsledku je homogénna vzorka proteínu. Medzi ďalšie faktory nutné k uspokojivej kryštalizácii patria: pH blízke izoelektrickému bodu, nízke teploty, vhodné rozpúšťadlo a pridané ióny či ligandy. Do vysoko saturovaného roztoku proteínu sa ďalej pridávajú rôzne soli, napr. síran amónny, ktoré znižujú rozpustnosť bielkoviny a umožňujú tak vytvoriť kryštalickú zrazeninu. [18]

Dostatočne vykryštalizovaný proteín sa umiestňuje do pôsobenia primárneho zväzku žiarenia. X-žiarenie s vlnovou dĺžkou spravidla 1,54 Å je generované buď prostredníctvom rotujúcej anódy rentgenky, v prípade malých kryštálov sa využíva synchrotron, ktorý dokáže vytvárať mnohonásobne väčšiu intenzitu žiarenia. Časť zväzku prechádza kryštálom nezmenene, časť je rozptýlená. Jednotlivé vlny rekombinujú a interferujú, celková difrakcia je potom zachytená na filme ako miera zčernania emulzie v závislosti na intezite rozptýleného žiarenia. V súčasnosti býva RTG film poväčšine nahradený CCD detektorom. Posudzuje sa difrakčný vzor, konkrétne ampitúdy a pozície jednotlivých rozptýlených lúčov. Keďže na úrovni atómov sa zväzok rozptyluje na elektrónoch, amplitúda rozptýlených vĺn udáva počet elektrónov a tým aj druh atómu. Konečným výstupom je mapa elektrónovej hustoty, ktorá sa vytvára analýzou intenzít a uhlov jednotlivých difrakcií. Informácie o intenzitách a uhloch získané priamo z RTG experimentu však pre zostrojenie mapy elektrónovej hustoty nie sú postačujúce. K určeniu pozície jednotlivých atómov a rekonštrukcii štruktúry potrebujeme navyše poznať fázu jednotlivých roptýlených vĺn. Jedná sa o tzv. fázový problém,

pretože fázové informácie nie je možné získať priamo z RTG experimentu. Fázový problém je riešiteľný buď výpočtovo alebo experimentálne, napr. s využitím metódy izomorfnej zámeny („Multiple Isomorphous Replacement“). Funguje na princípe vpravenia atómov ťažkých kovov do proteínu za predpokladu, že nezmenia jeho štruktúru. Fázovú informáciu získame z rozdielu v intenzitách obrazca pôvodnej a upravenej molekuly zostrojením Pattersonovej mapy. [6], [17], [18], [19]



Obr. 6: Vykryštalizovaný proteín [25]

Difrakčný vzor predstavuje veľké množstvo bodov s danými intenzitami, pozíciami a fázami jednotlivých vĺn odkazujúcich na usporiadanie molekúl v kryštále. Aplikovaním inverznej Fourierovej transformácie vznikajú mapy distribúcie elektrónovej hustoty molekúl. Skladaním jednotlivých vĺn so známou fázou v trojrozmernom priestore pomocou rotácie kryštálu získavame priestorový model týchto distribúcií. Ďalšou dôležitou úlohou je priradenie správneho reťazca k správnym dátam mapy elektrónovej hustoty („stopovanie reťazca“), čo je vo veľkej miere závislé na rozlíšení dát. Posledná fáza zdokonaľovania je iteratívny proces, ktorý sa snaží minimalizovať tzv. R-faktor. Ide v podstate o rozdiel medzi pozorovanými amplitúdami difrakcie a amplitúdami modelovo vypočítanými. Model je vyhodnotený ako úspešný v prípade, že sa R-faktor pohybuje do hodnoty 0,20. [20]

3.1.2 NMR spektroskopia

Spektroskopia s využitím nukleárnej magnetickej rezonancie sa oproti iným metódam určovania priestorovej štruktúry odlišuje najmä tým, že podkladové dáta môžu byť zaznamenané v roztoku, nie všetky proteíny totiž kryštalizujú ľahko. Roztoky sú prirodzeným prostredím proteínových makromolekúl. Naviac sa NMR vyznačuje schopnosťou zachytiť dynamické vlastnosti

štruktúr, ako napríklad termodynamické či kinetické aspekty interakcie medzi jednotlivými proteínmi či inými molekulami. [16]

Narozdiel od rentgenovej kryštalografie, NMR spektroskopia je obmedzená veľkosťami vzoriek, je vhodná pre vzorky o molekulovej hmotnosti menšej ako 50 kDa [6]. Metóda je založená na fakte, že jadrá určitých atómov (H, C, N, P) vykazujú spin, tj. magnetický moment. Na základe neho je možné skúmať chemické okolie týchto atómov a zároveň zisťovať vzdialenosti medzi jednotlivými atómami. Využívaný je predovšetkým vodíkový protón, pretože jeho výskyt je najčastejší. Pokiaľ sú takéto atómy, resp. ich jadrá, vystavené účinkom vonkajšieho magnetického poľa, ich spin sa zarovná v jednom smere pozdĺž poľa. Atóm sa z tohoto základného stavu môže dostať do stavu excitovaného aplikovaním impulzu elektromagnetického žiarenia. Ide o rádiofrekvenčný (RF) pulz, pričom dodaná energia pulzu musí zodpovedať rozdielu medzi energetickými hladinami. Jedná sa o frekvenciu, na ktorej dochádza k rezonancii. Pri návrate do základného stavu potom atóm emituje príslušnú RF, ktorú dokážeme zmerať. Táto frekvencia nám zároveň podáva informácie o okolí jadra, kde elektróny vytvárajú určité lokálne magnetické polia pôsobiace proti vonkajšiemu poľu a menia tak jeho účinok. Spôsobí to, že atómy potom rezonujú pri inej intenzite poľa, resp. iných RF. Jednotlivé líšiace sa frekvencie rovnakých jadier sa v spektrách označujú ako chemické posuny δ („chemical shifts“). Posuny sú udávané vzhľadom k referenčnej vzorke. Často však takáto jednodimenzionálna NMR nie je vhodná, pretože rozdiely v signáloch chemických posunov sú malé a prekrývajú sa.

V praxi sa preto využívajú prevažne 2D a 3D NMR experimenty. V rámci dvojdimenzionálneho spektra predstavuje diagonála klasické 1D spektrum, teda dáta vzťahujúce sa k rovnakému atómu. Píky mimo diagonálu poukazujú na interakcie vodíkových atómov, ktoré sú blízko seba v priestore. Menením hodnoty RF a jeho aplikáciou na jednotlivé H atómy dokážeme na základe týchto píkov odhaliť rôzne interakcie medzi nimi.

Existujú viaceré experimentálne prístupy v rámci NMR, medzi najznámejšie patria COSY a NOE. Pomocou COSY („Correlation Spectroscopy“) získavame informáciu o kovalentne spojených vodíkových atómoch. NOE spektrum („Nucelar Overhauser Effect“) spája páry vodíkových atómov, ktoré sú navzájom blízko seba v priestore, bez ohľadu na ich polohu v primárnej aminokyselinovej sekvencii. Zozbieraním a spracovaním týchto signálov dokážeme získať kompletnú informáciu o sekundárnej a terciárnej štruktúre. Miera vzájomných interakcií jadier je úmerná šiestej mocnine vzdialenosti. Tým pádom je možné definovať vzdialenosť jednotlivých vodíkových atómov v rámci reťazca. 2D spektrá sú vyhodnocované metódou sekvenčného priradenia („sequence assignment“), ktorá každému reziduu priraďuje tzv. „fingerprint“. Ide o jedinečnú kombináciu píkov v COSY spektre, pretože každé reziduum má špecifický súbor kovalentných spojení vodíkových atómov. Metóda teda umožnuje identifikáciu H atómov jednotlivých aminokyselín a zároveň je schopná určovať postranné reťazce rezidua. Z COSY spektier však nedostávame informáciu, ktorému reziduu v poradí aminokyselín daný vodík patrí, tj. na ktorej pozícii v reťazci sa nachádza. To môžeme odvodiť práve z NOE spektra, signalizujúceho susedné atómy v priestore.

Metódy využívajúce magnetizáciu H atómov sú vhodné pre štúdium menších proteínov, do väčších molekúl musia byt vpravné C alebo N atómy. [8], [19]

3.2 Výpočtové metódy

Jednou z najvýznamnejších úloh bioinformatiky je určenie priestorovej konformácie proteínu *ab initio*, od počiatku. Znamená to snahu zrekonštruovať terciárnu, prípadne kvartérnu štruktúru proteínu priamo zo známej aminokyselinovej sekvencie. Kľúčovým krokom v tomto procese je práve odhad sekundárnej štruktúry proteínu z jeho primárnej štruktúry. Výsledky odhadu 2D štruktúry sú nezanedbateľne nápomocné pri predikcii bodových mutácií, identifikovaní proteínových tried, predikovaní epitopov a celkovo dizajnovaní nových proteínov. V oblasti predikcie sekunárnej štruktúry rozlišujeme dva hlavné smery:

1. predikcie štruktúrnej triedy (tzv. „all α-“, „all β-“, α/β triedy)

2. predikcie štruktúrnuch stavov pre každé jednotlivé reziduum proteínu. Metódy rozobrané v nasledujúcich podkapitolách riešia druhý spôsob predikcie. [27]

3.2.1 Metóda Chou-Fasman

Chou-Fasmanov algoritmus sa radí k základným metódam predikcie sekundárnej štruktúry. Podkladom jeho vzniku bola databáza experimentálne zistených proteínových štruktúr. Na ich základe sa stanovilo početné zastúpenie určitej aminokyseliny v jednotlivých formách sekundárnej štrukúry. Metóda je schopná detekovať štyri formy štruktúrneho usporiadania: α-hélix, β - skladaný list, β-ohyb, prípadne žiadnu z uvedených, náhodný zhluk („coil“). Výsledkom štatistického zhodnotenia dát bolo stanovenie preferenčných parametrov k jednotlivým druhom sekundárnych štruktúr (Tab. 2):

$$P(S) = P\frac{(S|R)}{P(R)} \quad , \tag{1}$$

Preferenčný parameter *P* pre štruktúrnu formu *S* je pomer pravdepodobnosti výskytu daného rezidua *R* vo forme *S* a pravdepodobnosti výskytu rezidua *R* v celej databáze. Parametre teda určujú tendencie jednotlivých reziduí vytvárať danú štruktúrnu formu. Podľa týchto hodnôt sa reziduám prideľujú označenia:

- „former“ (P > 1) – silný („H“), slabý („h“)
- „breaker“ (P < 1) – silný („B“), slabý („b“)
- indiferentná aminokyselina (P = 1) – „I“, „i“

Pre β- ohyb je zároveň určená pozičná preferencia. Niektoré reziduá sa totiž vyskytujú na určitej pozícii s väčšou pravdepodobnosťou ako na inej (prolín) [10], [11].

Parametre tvoria základ empirických predikčných pravidiel:

1. Každej aminokyseline sa priradia preferenčné hodnoty na základe tabuľky pozdĺž celého polypeptidového reťazca.

2. Vytvorí sa skenovacie okno o dĺžke šesť aminokyselinových reziduí. Začiatočný úsek α- helixu indikuje zhluk štyroch po sebe nasledujúcich aminokyselín s preferenčnými hodnotami patriacimi do skupiny „former“ (Hα, hα). Nájdený helikálny úsek (jadro hélixu) je rozširovaný po oboch stranách reťazca, kým sa nenarazí na tetrapeptidovú oblasť s reziduami „breaker“ (Bα, bα). Celkovo musí byť pre hélix splnená podmienka, že priemerná hodnota P (α) > 1,03 a zároveň P (α) > P (β). Týmto spôsobom sú vymedzené helikálne úseky. Rozšírené verzie uvažujú, že niektoré reziduá typicky tvoria štruktúry v určitých oblastiach reťazca.
3. Vytvorí sa skenovacie okno o dĺžke päť aminokyselinových reziduí. Jadro β-skladaného listu predstavuje zhluk troch po sebe nasledujúcich aminokyselín s preferenčnou hodnotou P (β) > 1 (Hβ, resp. hβ). V prípade, že sa objaví tetrapeptidová oblasť reziduí Bβ, prípadne bβ (P (β) < 1), jedná sa o ukončenie β-listu. Pre β-list musí byť ďalej splnená podmienka, že celková tendencia úseku P (β) > 1,05 a zároveň P (β) > P (α)
4. U prekrývajúcich sa α- a β-úsekov sa výsledná štruktúra určí na základe väčšej z priemerných hodnôt P (α) a P (β) pre danú oblasť.
5. Úseky β-ohybu sú určené podľa vypočítaného parametru p(t), kde:

$$p(t) = f(i) \cdot f(i+1) \cdot f(i+2) \cdot f(i+3) \quad , \qquad (2)$$

pričom *i* indexuje polohu daného rezidua. V prípade, že $p(t) > 7.5 \cdot 10^{-4}$, daná oblasť má tendenciu vytvárať β-ohyb. Aby bol úsek skutočne predikovaný ako ohyb, musí byť zároveň splnená podmienka, že priemerná hodnota P (turn) > 1 pre štyri po sebe nasledujúce reziduá a zárveň P (α) < P (turn) > P (β).

6. V prípade, že výsledný predikovaný úsek je kratší ako jadro definované pre danú štruktúru, sa táto predikcia anuluje.
7. Úseky, ktorým nebola priradená žiadna z troch štruktúr, sú označené ako náhodné zhluky. [12], [13]

Tab. 2: Preferenčné hodnoty jednotlivých aminokyselín vzhľadom k sekundárnej štruktúre [11]

α-helix			β-skladaný list			β-ohyb					
Reziduum	**P(α)**	**Ozn.**	**Reziduum**	**P(β)**	**Ozn.**	**Reziduum**	**P(turn)**	**f(i)**	**f(i+1)**	**f(i+2)**	**f(i+3)**
Glu	1,44	Hα	Val	1,64	Hβ	Asn	1,56	0,161	0,083	0,191	0,091
Ala	1,39	Hα	Ile	1,57	Hβ	Gly	1,56	0,102	0,085	0,190	0,152
Met	1,32	Hα	Thr	1,33	hβ	Pro	1,52	0,102	0,301	0,034	0,068
Leu	1,3	Hα	Tyr	1,31	hβ	Asp	1,46	0,147	0,110	0,179	0,081
Lys	1,21	hα	Trp	1,24	hβ	Ser	1,43	0,120	0,139	0,125	0,106
His	1,12	hα	Phe	1,23	hβ	Cys	1,19	0,149	0,050	0,117	0,128
Gln	1,12	hα	Leu	1,17	hβ	Tyr	1,14	0,082	0,065	0,114	0,125
Phe	1,11	hα	Cys	1,07	hβ	Lys	1,01	0,055	0,115	0,072	0,095
Asp	1,06	hα	Met	1,01	Iβ	Gln	0,98	0,074	0,098	0,037	0,098
Trp	1,03	Iα	Gln	1,00	Iβ	Thr	0,96	0,086	0,108	0,065	0,079
Arg	1,00	Iα	Ser	0,94	iβ	Trp	0,96	0,077	0,013	0,064	0,167
Ile	0,99	iα	Arg	0,94	iβ	Arg	0,95	0,070	0,106	0,099	0,085
Val	0,97	iα	Gly	0,87	iβ	His	0,95	0,140	0,047	0,093	0,054
Cys	0,95	iα	His	0,83	iβ	Glu	0,74	0,056	0,060	0,077	0,064
Thr	0,78	iα	Ala	0,79	iβ	Ala	0,66	0,060	0,076	0,035	0,058
Asn	0,78	iα	Lys	0,73	bβ	Met	0,6	0,068	0,082	0,014	0,055
Tyr	0,73	bα	Asp	0,66	bβ	Phe	0,6	0,059	0,041	0,065	0,065
Ser	0,72	bα	Asn	0,66	bβ	Leu	0,59	0,061	0,025	0,036	0,070
Gly	0,63	Bα	Pro	0,62	Bβ	Val	0,5	0,062	0,048	0,028	0,053
Pro	0,55	Bα	Glu	0,51	Bβ	Ile	0,47	0,043	0,034	0,013	0,056

3.2.2 GOR (Garnier-Osguthorpe-Robson) metóda

V porovnaní s Chou-Fasmanovou metódou, ktorá uvažuje pri predikcii iba jednu samostatnú aminokyselinu ako faktor rozhodujúci o sekundárnej štruktúre, metóda GOR prichádza s určitým zdokonalením. Predpokladá vplyv okolitých reziduí, čím sa zvyšuje presnosť metódy. Celkovo je teda pri predikcii určitého rezidua uvažovaných 17 reziduí, 8 reziduí pred a 8 reziduí za touto pozíciou. GOR predstavuje štatistickú metóda vychádzajúca z Bayesovskej podmienenej pravdepodobnosti a informačnej teórie. Základná myšlienka metódy sa odvíja od informačnej funkcie v tvare:

$$I(S;R)=\log[P(S|R)/P(S)] \quad , \tag{3}$$

kde S je jedna zo štruktúrnych foriem, R je dané reziduum a P $(S|R)$ podmienená pravdepdobnosť výskytu S štruktúrnej formy, ak je prítomné R reziduum. [9]

Zároveň platí (z Bayesovskej pravdepodobnosti):

$$P(S|R)=P(S,R)/P(R) \quad , \tag{4}$$

kde P (S, R) je spoločná pravdepdobnosť výskytu R rezidua a zároveň výskytu S štruktury a $P(R)$ pravdepdobnosť výskytu R rezidua.

Pravdepodobnosti môžeme vyjadriť vo forme frekvencií výskytu reziduí a štruktúr v známej databáze:

$$I(S;R)=\log(f_{S,R}/f_S) \ , \tag{5}$$

kde $f_{S,R}$ je frekvencia výskytu rezidua R v štruktúre S a f_S frekvencia výskytu všetkých reziduí v štruktúre S.

Potom môžeme vypočítať tzv. informačné rozdiely medzi jednotlivými reziduami ovplyvňujúcimi predikované reziduum na pozícii m:

$$I(\Delta Sm;R_1,..R_x)=\log\frac{P(Sm,R_{1,}..R_x)}{1-P(Sm,R_{1,}..R_x)}+\log\frac{1-P(S)}{P(S)} \ , \tag{6}$$

kde ΔS predstavuje informačný rozdiel medzi hypotézou, že R sa vyskytuje v konformácii S a hypotézou, že R sa v tejto konformácii nevyskytuje (je v inej konformácii). Môžme ho teda definovať ako rozdiel informačných funkcií *I (S; R)* a *I (nonS; R)*. Z tejto rovnice dokážeme vyjadriť podiel združených pravdepodobností :

$$\frac{P(Sm,R_1,..Rx)}{1-P(Sm,R_{1,}..Rx)}=\frac{P(S)}{1-P(S)}\cdot e^{-I(\Delta Sm;R_1,..Rx)} \ . \tag{7}$$

Po aproximovaní pravdepodobnosti ako frekvencie výskytu jednotlivých stavov, sa tieto podiely využívajú ako hodnoty skórovací matíc pre predikciu štruktúry. [13], [14]

Samotný algoritmus predikcie pozostáva zo zostrojenia štyroch skórovacích matíc o rozmeroch 20x17. Každej zo 17 pozícii v polypeptidovom reťazci, ktoré berieme do úvahy, zodpovedá stĺpec s hodnotami parametrov pre každú z 20 možných aminokyselín. Jednotlivé skórovacie matice reprezentujú hodnoty pre jednu zo štyroch foriem štruktúry, ktorú môže centrálne reziduum zaujať. (hélix, list, ohyb, náhodný zhluk). Hodnoty pravdepodobností v každej matici sa vypočítajú na základe vzťahov uvedených vyššie pre každú aminokyselinu a každú pozíciu. Následne sa prechádza požadovanú sekvencia skenovacím oknom o dĺžke 17 reziduí a na základe matíc sa vypočíta celkové skóre pre centrálne reziduum ako suma pravdepodobností všetkých pozícií. Štruktúrna forma sa priraďuje podľa matice, v rámci ktorej sa dosiahne najvyššie skóre. [15]

Jednotlivé metódy GOR sa postupne vyvíjali. Kvôli zjednodušeniu GOR I a GOR II zaviedli predpoklad, že každé zvažované reziduum v okolí centrálneho rezidua naň pôsobí samostatne, bez vzájomných interakcií s ostatnými reziduami. Metóda GOR III už predpokladá párový vplyv okolitých reziduí, teda určitú koreláciu medzi reziduami. Najnovšia metóda GOR V zvyšuje presnosť predikcií s využítím viacnásobného zarovnania. [9], [13], [14]

3.2.3 Metóda najbližšieho suseda („k-Nearest Neighbor“)

Na rozdiel od predchádzajúcich metód, „k-Nearest Neighbor“ metóda neposkytuje priradenie k jednotlivým štruktúrnym stavom, ale skôr pravdepodobnostné rozloženie nad týmito stavmi. Základná myšlienka spočíva v nájdení určitého množstva *k*-najbližších susedov k predikovanej sekvencii. Jedná sa o sekvencie s podobnou primárnou štruktúrou (homológne sekvencie) a známou sekundárnou štruktúrou prístupnou z databázy. Predpokladá sa, že sekvencie s podobným obsahom a sledom aminokyselín v reťazci budú taktiež vytvárať podobné sekundárne štruktúry. Predikcia prebieha na princípe priradenia príslušnosti k štruktúrnej triede pre centrálne reziduum zo zvoleného okna. Najbližší susedia sa vyberajú z databázy sekvencií s aplikovaním okna rovnakej dĺžky. [13]

Pre proteíny databázy a predikovaný proteín sa vytvorí skórovacia matica, najčastejšie PSSP („Position Specific Scoring Matrix“), ktorá predstavuje vstup predikčného systému. Následne sa získa sekvenčný profil jednotlivých reziduí z viacnásobného zarovnania sekvencií aplikovaním algoritmu PSI-BLAST. Ten je potrebný k výpočtu vzdialenostnej miery *d,* ktorá vyčleňuje najbližších susedov sekvencie:

$$d = max\{1, (\sum_{i=1}^{20}\sum_{j=1}^{W}\left|(p_{ij}^{\ sekv} - p_{ij}^{datab})\, min\{j, (W+1-j)\}\right|)\} \quad , \tag{8}$$

kde W je veľkosť okna, p_{ij} profil *j*-tej pozície v okne pre *i*-tu aminokyselinu skúmanej sekvencie a homológnej sekvencie z databázy. Vzdialenostná miera je pozične váhovaná, berie teda do úvahy fakt, že vzdialenejšie reziduá majú na štruktúru centrálneho rezidua okna menší vplyv ako reziduá v jeho blízkosti.

Ďalším krokom metódy je priradenie príslušnosti vybraných susedov k štruktúrnej triede. Využíva sa fragment s oknom rovnakej dĺžky ako u predikovanej sekvencie. Pôvodné „Nearest Neighbor“ metódy priraďovali príslušnosť priamo podľa štruktúry centrálneho rezidua. Hodnoty príslušnosti k triedam boli teda výlučne binárne. Novšie metódy zohľadňujú vplyv okolitých reziduí na štruktúrnu formu centrálneho rezidua. Berie sa tiež ohľad na vzdialenosť reziduií od centrálnej pozície, na základe čoho sú okolité reziduá váhované. Váhy sa väčšinou určujú ako prevrátená hodnota vzdialenosti od centrálnej pozície, keďže so vzdialenosťou klesá vplyv na štruktúrnu formu vybraného rezidua. Príslušnosť k štruktúrnej triede je susedom priradená zväčša ako percento výskytu susedov (reziduí) v danej triede váhované uvedenými váhami.

U novších metód má priradenie štruktúrnej triedy samotnému predikovanému reziduu fuzzy formu. Miera príslušnosti k triede sa pohybuje v rozmedzi <0,1>, kde 0 indikuje, že reziduum nepatri do danej triedy a 1 indikuje úplnú príslušnosť k triede. Príslušnosť $u_i\,(r)$ rezidua $\boldsymbol{r}$ ku každej z tried i možno vyjadriť vzťahom:

$$u_i(r) = \frac{\sum_{j=1}^{K} u_{ij}\left(1/\left|d\left(r, r_j\right)^{2/m-1}\right|\right)}{\sum_{j=1}^{K}\left(1/\left|d\left(r, r_j\right)^{2/m-1}\right|\right)} \quad . \tag{9}$$

Podľa vzťahu je príslušnosť k triede daná počtom susedov priradených k tejto triede u_{ij} a vzdialenostnou mierou $d\ (r,r_j)$ medzi predikovaným reziduom r a príslušným reziduom suseda r_j. Premenná m, vystupujúca vo vzťahu predsatvuje tzv. fuzzifikátor. Fuzzifikátor určuje, ako je hodnota príslušnosti k triede závislá na vzdialenosti medzi reziduami. Konkrétny spôsob priraďovania sa u jednotlivých algoritmov líši a má výrazný vplyv na účinnosť algoritmu. Zavedenie fuzzy klasifikácie do metódy „Nearest Neighbor“ má značný význam, pretože umožnuje určiť mieru istoty s akou patrí skúmané reziduum k jednotlivým štruktúrnym triedam. [30]

Dôležitým krokom v predikcii je vhodný výber parametrov, ako je počet susedov k ,zohľadňovaných v predikcii a dĺžka okna. Pôvodne sa volil parameter k v rozmedzí 1-25, empiricky sa však zistilo, že najlepších výsledkov sa dosahuje pri k v rozmedzí 50-200 . Paradoxne, dĺžka okna nie je úmerná kvalite predikcie, napriek tomu, že určitým zväčšovaním okna je môžné lepšie zachytiť globálne interakcie. Algoritmus NN dosahuje najvyššej presnosti predikcie pri zvažovaní 9 reziduí pred a 9 reziduí za centrálnym reziduom, teda s použitím okna o dĺžke 19 reziduí. [29]

3.2.4 PhD

Metóda PhD využíva viacnásobné zarovnanie sekvencií (MSA) v kombinácií s neurónovými sieťami. Z viacnásobného zarovnania je vytvorený sekvenčný profil, tj. zastúpenie jednotlivých reziduí v jednotlivých stĺpcoch zarovnania. Architektúra metódy je navrhnutá do troch úrovní:

1. Neurónová sieť sekvencia → štruktúra

   Vstupom dvojvrstvovej doprednej siete sú profily v rámci okna s dĺžkou 13 reziduí ($a_{j-6}...a_j...a_{j+6}$), centrálnu pozíciu predstavuje reziduum a_j , ktorého štruktúru predikujeme. Každé reziduum (resp. pozícia okna) teda produkuje 20 vstupov, jeden vstup je vymedzený pre pre prípad, že okno presahuje C- alebo N-koniec proteínu a ďalší udáva kvalitu MSA zarovnania. Sieť následne pozostáva z 13x22 uzlov. Výstupom siete sú 3 váhy – pravdepodobnostné hodnoty príslušnosti predikovaného rezidua k jednotlivým štruktúrnym triedam: $P(j\alpha)$ – hélix, $P(j\beta)$ – skladaný list a. $P(jC)$ – náhodný zhluk

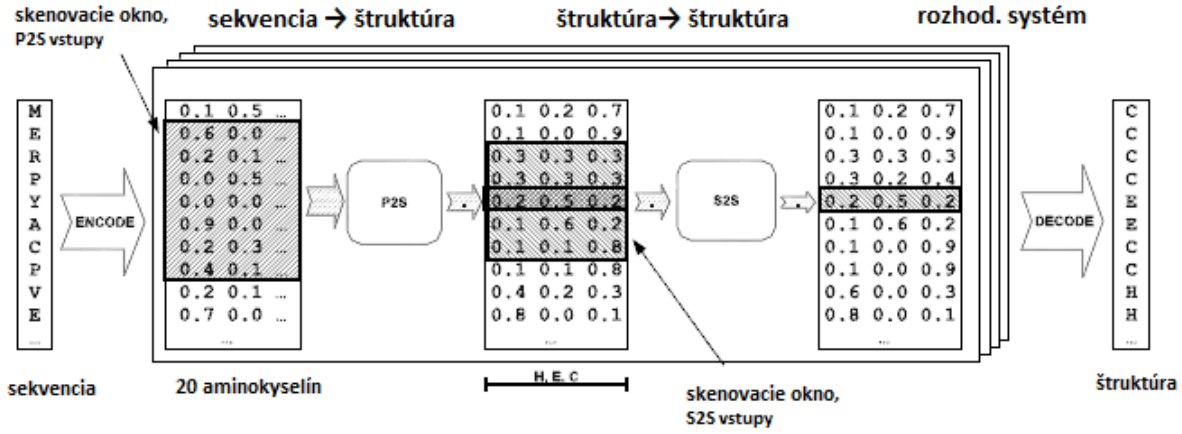
2. Neurónová sieť štruktúra → štruktúra

   Vstup siete predstavuje okno o dĺžke 17 reziduí, pričom každé reziduum je reprezentované tromi váhami z výstupu prvého stupňa sekvencia → štruktúra. Jedna váha je opäť vyhradená pre prípad, že okno presahuje jeden z koncov proteínu. Sieť je teda tvorená

17x4 jednotkami a výstupom sú upravené pravdepodobnosti možných štruktúrnych stavov.

3. Rozhodovací systém

Keďže sú neurónové siete závislé na rade paramatrov – výber trénovacích dát, poradie trénovania atď., PhD využíva väčší počet separátne trénovaných párov sietí. Jednotlivé predikcie sú privádzané na vstup tretej siete, označovanej ako rozhodovací systém. Na tejto úrovni dochádza k spriemerovaniu jednotlivých predikcií a výsledný výstup predstavuje štruktúrna forma s najväčším priemerným skóre. Zároveň sú anulované predikcie krátkych štruktúrnych úsekov, ktoré sú následne označené ako „náhodný zhluk“. [27], [28], [32]



Obr. 7: Architektúra siete – metóda PhD [31]

4 Praktická realizácia predikčného programu

Pre samotnú realizáciu predikcie bola zvolená Chou-Fasmanova metóda, ktorá je vďaka svojej jednoduchosti a rýchlosti spracovania častou voľbou pre orientačné stanovovanie štruktúry. Chou- Fasman predstavuje vo všeobecnosti základnú metódu predikcie sekundárnej štruktúry, na ktorú nadväzujú niektoré zložitejšie metódy, prípadne ju určitým spôsobom rozširujú. Zároveň je metóda uskutočňovaná po krokoch a tým pádom pomerne dobre názorná. U mnohých iných metód, využívajúcich napr. umelé neurónové siete, väčšina rozhodovacích procesov prebieha skryte, resp. neposkytuje adekvátny popis svojich rozhodnutí užívateľovi. Chou - Fasmanovu metódu teda možno z uvedených dôvodov považovať za vhodnú k tomuto účelu.

Program, vrátane grafického užívateľského rozhrania, je realizovaný v programovom prostredí Matlab R2009/a, ktorý je vďaka bioinformatickým toolboxom dobre využiteľný pre prácu s biologickými sekvenciami.

4.1 Výber dát

Vstupnými dátami každého predikčného algoritmu sekundárnej štruktúry sú sekvencie aminokyselín v jednopísmenovom kódovaní (Tab. 1). Realizovaný program, podobne ako väčšina predikčných softvérov, pracuje so súbormi vo formáte FASTA, ktorý zároveň podporuje väčšina proteínových databází. Tento textovo založený formát obsahuje hlavičku s jednoriadkovým popisom proteínovej sekvencie a samotnú znakovú reprezentáciu sekvencie. Tieto časti sú odlíšené identifikátorom „>“, ktorý uvádza popisný riadok sekvencie.

Pre testovanie vytvoreného programu a následne voľne dostupných internetových softvérov bol zvolený nasledujúci súbor významných ľudských proteínov:

1. HBB [Human] – hemoglobín (podjednotka β)
2. CRP [Human] – C-reaktívny proteín
3. ACTS [Human] – aktín α
4. ALBU [Human] – albumín
5. P53 [Human] – bunkový tumorový antigén p53

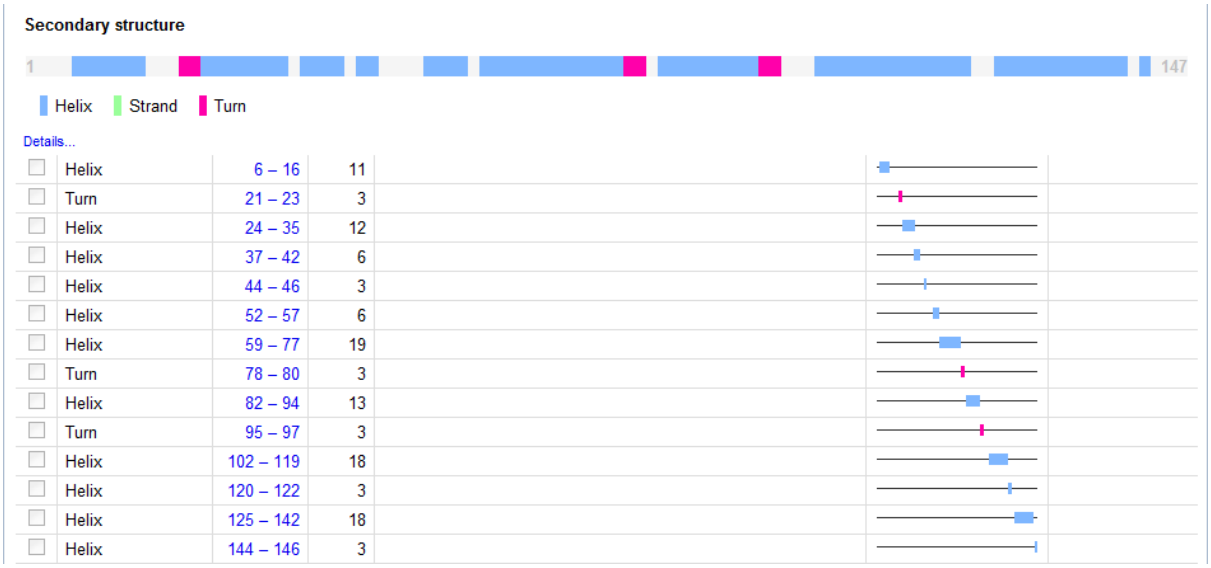
4.1.1 Databáza UNIPROT

Dáta pre testovanie samotného programu a porovnávanie vybraných softvérov pochádzajú z UNIPROT databázy. Ide o jeden z najkomplexnejších zdrojov proteínových sekvencií a dát súvisiacich s proteínmi. Keďže sa zároveň jedná o databázu známych sekundárnych štruktúr, pochádzajúcich prevažne z RTG a NMR experimentov, pri porovnávaní výsledkov predikcií ju môžno považovať za referenčnú.

Nasledujúca časť je prehľadom reprezentácie zvolených dát (sekvencií vo FASTA formáte a ich známych sekundárnych štruktúr) v databáze UNIPROT, dostupnej z http://www.uniprot.org/. V zátvorke je uvedené identifikačné číslo sekvencie v rámci databázy.

Sekvencia 1: HBB (P68871)

```
>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens GN=HBB
PE=1 SV=2
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

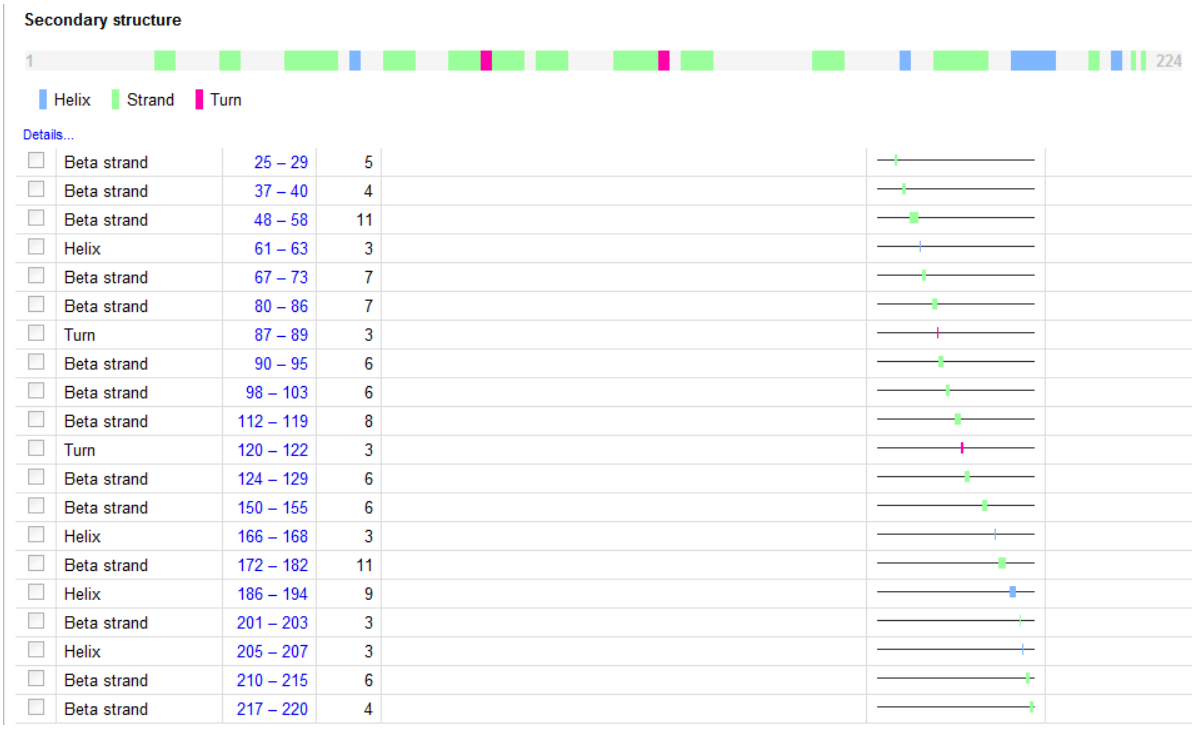


☐	Helix	6 – 16	11
☐	Turn	21 – 23	3
☐	Helix	24 – 35	12
☐	Helix	37 – 42	6
☐	Helix	44 – 46	3
☐	Helix	52 – 57	6
☐	Helix	59 – 77	19
☐	Turn	78 – 80	3
☐	Helix	82 – 94	13
☐	Turn	95 – 97	3
☐	Helix	102 – 119	18
☐	Helix	120 – 122	3
☐	Helix	125 – 142	18
☐	Helix	144 – 146	3

Obr. 8: Sekundárna štruktúra sekvencie 1 z databázy UNIPROT

Sekvencia 2: CRP (P02741)

```
>sp|P02741|CRP_HUMAN C-reactive protein OS=Homo sapiens GN=CRP PE=1
SV=1
MEKLLCFLVLTSLSHAFGQTDMSRKAFVFPKESDTSYVSLKAPLTKPLKAFTVCLHFYTE
LSSTRGYSIFSYATKRQDNEILIFWSKDIGYSFTVGGSEILFEVPEVTVAPVHICTSWES
ASGIVEFWVDGKPRVRKSLKKGYTVGAEASIILGQEQDSFGGNFEGSQSLVGDIGNVNMW
DFVLSPDEINTIYLGGPFSPNVLNWRALKYEVQGEVFTKPQLWP
```



Obr. 9: Sekundárna štruktúra pre sekvenciu 2 z databázy UNIPROT

Sekvencia 3: ACTS (P68133)

```
>sp|P68133|ACTS_HUMAN Actin, alpha skeletal muscle OS=Homo sapiens GN=ACTA1
PE=1 SV=1
MCDEDETTALVCDNGSGLVKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEA
QSKRGILTLKYPIEHGIITNWDDMEKIWHHTFYNELRVAPEEHPTLLTEAPLNPKANREK
MTQIMFETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRL
DLAGRDLTDYLMKILTERGYSFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSSLEK
SYELPDGQVITIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNV
MSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWIT
KQEYDEAGPSIVHRKCF
```

Obr. 10: Sekundárna štruktúra pre sekvenciu 3

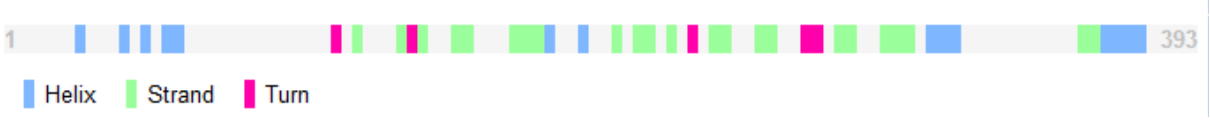
Sekvencia 4: ALBU (P02768)

```
>sp|P02768|ALBU_HUMAN Serum albumin OS=Homo sapiens GN=ALB PE=1 SV=2
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPF
EDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCTVATLRETYGEMADCCAKQEP
ERNECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLF
FAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAV
ARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLK
ECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYAR
RHPDYSVVLLLRLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFE
QLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVV
LNQLCVLHEKTPVSDRVTKCCTESLVNRRPCFSALEVDETYVPKEFNAETFTFHADICTL
SEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLV
AASQAALGL
```

Obr. 11: Sekundárna štruktúra pre sekvenciu 4

Sekvencia 5: p53 (P04637)

```
>sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens GN=TP53
PE=1 SV=4
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```



Obr. 12: Sekundárna štruktúra pre sekvenciu 5

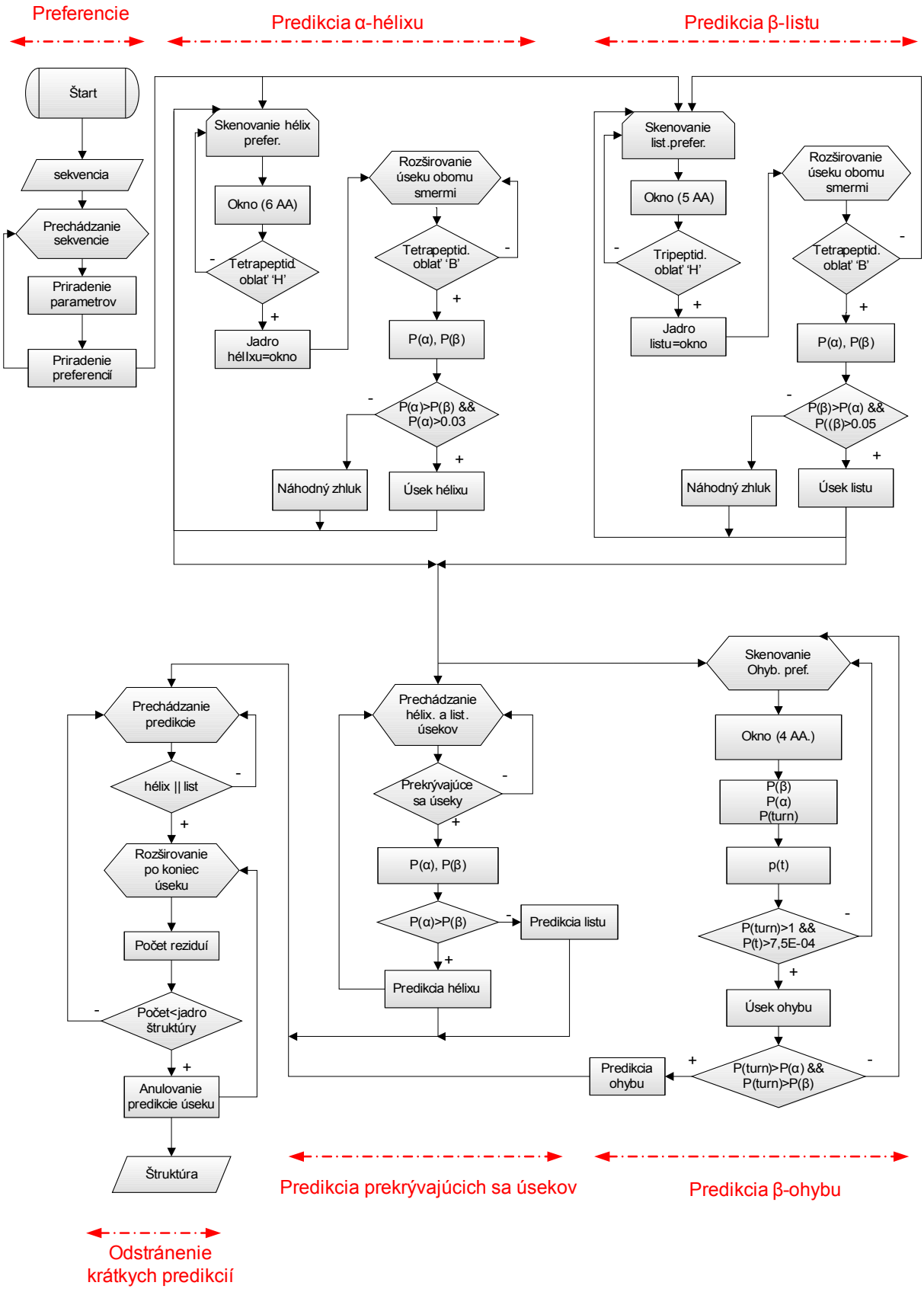
4.1.2 Požadované výstupy

Požadovaným výstupom predikcie je sekvencia transformovaná do postupnosti štruktúrnych tried podľa štandardného kódovania DSSP („Database of Protein Structure Assignment“). Ide o databázu jednotlivých štruktúrnych priradení obsahujúcu osem základných tried sekundárnej štruktúry: H („Helix“, α-hélix), E („Extended“, β-skladaný list), G (3_{10}-hélix), I (π-hélix), B („Bridge“, jednoreziduový β-list), T („Turn“, β-ohyb), S („Bend“), C („Coil“, náhodný zhluk nezaradený do žiadnej z tried). Vo väčšine prípadov sú predikované štruktúry redukované na tri triedy, v prípade nasledujúceho algoritmu je využitá štvorstavová redukcia podľa tzv. CASP štandardu: {H, G, I} → H, {B, E} → E, {S, T} → T, {C} → C.

4.2 Návrh predikčného algoritmu

Algoritmus pre predikciu sekundárnej štruktúry je navrhnutý na základe Chou-Fasmanovej metódy a rozhodovacích pravidiel zo str. 16. Grafické znázornenie zjednodušeného návrhu poskytuje vývojový diagram na Obr. 13.

Ná základe diagramu možno predikčný algoritmus rozdeliť do piatich súborov blokov. V prvej časti sú vstupnej sekvencii priraďované preferencie k sekundárnym štruktúram. Predikcia sa následne vetví na predikciu hélixu a listu. Bloky obsahujúce výraz *P(x)* počítajú priemerné preferenčné hodnoty reziduí na danej oblasti pre daný druh štruktúry *x*. Rozhodovacie pravidlá a postup samotných predikcií znázornených v diagrame je uvedený v popise samotnej metódy. Výsledkom sú úseky hélixov a listov priradené príslušným pozíciam. V ďalšom súbore blokov sa úseky „zjednocujú“ do jednej postupnosti pre danú sekvenciu spracovaním prekrývajúcich sa oblastí. Samostatnú časť predstavuje vyčlenenie úsekov ohybu. Blok *p(t)* počíta pozičnú preferenciu pre potenciálny ohybový úsek. Pri predikcii ohybu sa prekrývajúce oblasti nevyhľadávajú, po splnení príslušných podmienok je úsek do výslednej predikcie dosadený automaticky. Čiastočné predikcie priradené pozíciam sa spájajú do jednej výslednej postupnosti predikcií v poslednom súbore blokov. Ten zabezpečuje anulovanie krátkych predikcií štruktúr s nedostatočným počtom reziduí, ktoré vznikli v dôsledku prekrývajúcich sa oblastí.



Obr. 13: Vývojový diagram realizovaného programu s vyznačenými funkčnými celkami

4.3 Popis programovej realizácie, ukážka funkcie

Predikciu sekundárnej štruktúry realizuje funkcia `ChouFasman`. Vstupom je sekvencia vo formáte FASTA, načítaná zo súboru pomocou funkcie `fastaread`, prípadne priamo zadaná sekvencia vo forme reťazca. Výstupom je sekundárna štruktúra vo forme uvedenej v podkapitole 4.1.2.

Funkcia v tvare `[hlavicka,sekvencia]=fastaread([PathName,FileName])` získava vstupné parametre o umiestnení súboru z výstupu funkcie `uigetfile`, ktorá otvára dialógové okno pre výber súboru s vymedzeným formátom FASTA. Príklad fungovania programu môže byť demonštrovaný na krátkej fiktívnej sekvencii:

```
sekvencia =
PEMMLMNFPPPPMFPCFFLLVLYYYYGDEEDQTDMSRKKESDTSYVYYYYPLTQTIDSFWRWCSPPPP
```

Program pracuje s preferenčnými parametrami pre jednotlivé štruktúrne triedy – *P(α)*, *P(β)* , *P (turn)* a *p(t)* z tabuľky č. 2, uloženými v štruktúre `Pref_param`. Každej aminokyseline sekvencie sú priradené hodnoty parametrov pre každú štruktúru vo forme vektorov `Val_helix`, `Val_list` a `Val_ohyb`. Pre predikciu ohybov sú reťazcu zároveň priradené pozičné preferencie pre možné polohy aminokyseliny v tetrapeptidovom ohybovom úseku – `pt1` až `pt4`. Pre predikciu hélixov a listov sú príslušné hodnoty porovnávané s empiricky stanovenými hranicami (Tab. 2) a každej aminokyseline reťazca je priradená miera preferencie k jednotlivým triedam v podobe znakov *H*, *h*, *B*, *b*, *I*, *i* . Týmto spôsobom vzniknú vektory preferencií k štruktúram – `Pref_helix` a `Pref_list`.

```
Pref_helix =
BHHHHHihBBBBHhBihhHHiHbbbbBhHHhhihHbIhhHbhibbibbbbBHihiihbhIIIibBBBB
Pref_list =
BBIIhIbhBBBBIhBhhhhhHhhhhhibBBbIhbIiibbBibhihHhhhhBhhIhHbihhihhiBBBB
```

Nasledujúcou časťou je prevod preferencií do postupnosti štruktúrnych tried. Každá uvažovaná trieda je predikovaná najprv samostatne a uložená vo forme vektorov `Strukt_l` (list), `Strukt_h` (helix) a `Strukt_o` (ohyb). Tieto vektory sú prealokované na indiferentnú triedu „C“. Preferenčné vektory `Pref_helix` a `Pref_list` sú skenované oknom príslušnej dĺžky pre konkrétny druh štruktúry. Funkcia `strcmpi` porovnáva znaky okna s tripeptidovým (list) alebo tetrapeptidovým (hélix) zhlukom „formerov“ v ľubovoľnom zložení („H“, „h“) a poradí. Okno, ktoré

obsahuje takýto zhluk, je vyhodnotené ako „jadro“ danej štruktúry. V príslušných oblastiach vektorov `Strukt_l`, resp. `Strukt_h` je zaznamená jeho poloha symbolom štruktúrnej triedy („E“, resp. „H“). Jadro sa rozširuje doplňovaním znakov do vektorov štruktúr, kým `strcmpi` nevyhodnotí tetrapeptidovú oblasť „breakerov“, taktiež v ľubovoľnom zložení a poradí. Cyklus rozširovania je v tomto prípade zastavený, analogicky sa rozširuje v opačnom smere. Takto sú vymedzené helikálne a listové úseky. Medzi začiatočnou a koncovou polohou úseku sú vypočítané priemerné preferencie *P(α)* a *P(β)* z vektorov `Val_helix` a `Val_list`. V prípade, že nie sú splnené podmienky metódy (viď príslušné bloky diagramu), je ich predikcia anulovaná (na úsek sú dosadené znaky „C“). Začína nové skenovanie preferenčných parametrov s oknom posunutým na nepredikovanú oblasť.

Pre určenie ohybu sa využíva tetrapeptidové okno, v rámci ktorého sa počíta priemerná preferenčná hodnota pre ohyb. Pokial je splnená podmienka pre hodnotu *P(turn)* na danom úseku a zároveň podmienka pre pozičné preferencie *p(t)*, na príslušnú tetrapeptidovú oblasť vektoru `Strukt_o` sú dosadené znaky „T“. Okno sa posúva na nepredikovanú oblasť.

```
Strukt_helix =
HHHHHHHHCCCCHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCC
Strukt_list =
CCCCCCCCCCCCEEEEEEEEEEEEEECCCCEEEEEEEEEEEEEEEEEEEEEEEEEEEEECCCC
Strukt_ohyb =
CCCCCCCCCCCCCCCCCCCCCTTTTCCCCTTTTCTTTTTTTTCCCTTTTCCCCCCTTTTCCCTTTTC
```

Čiastočné predikcie sú v ďalšom cykle prechádzané a zjednotené do výsledného vektoru štruktúry `Strukt`. Ohybové oblasti (vždy štvorice aminokyselín) sú do vektoru dosadené v prípade, že vypočítaná preferencia úsekov je väšia ako preferencia k listu a zároveň k hélixu. Prekrývajúce sa oblasti štruktúr listov a hélixov sa rieša nasledovne: na úseku, kde je predikovaná štruktúra pre obe triedy, sa vypočíta priemerná preferenčná hodnota reziduí k obom triedam. Úsek označený symbolom triedy je dosadený do výsledného vektoru, zvyšné pozície sú doplnené pôvodnými znakmi vektorov.

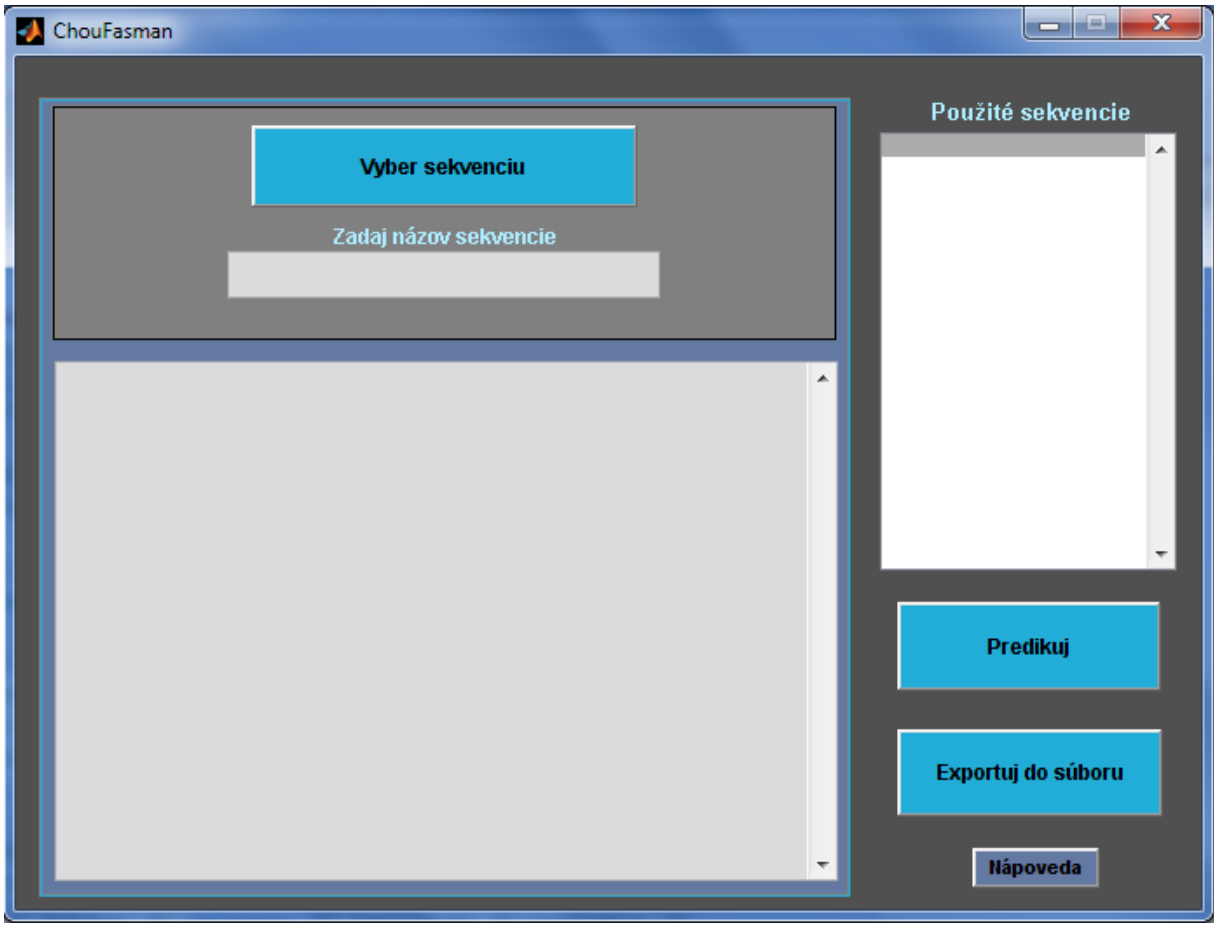
Posledný cyklus spresňuje predikciu vyhodnotením dĺžky jednotlivých úsekov vo výslednom reťazci a v prípade, že je predikovaná oblasť menšia ako jadro danej štruktúrnej triedy, predikciu anuluje. Výstupný vektor `Strukt` v daných pozíciach obsahuje znak „C“, rovnako ako v prípade, že danej pozícii nebola žiadna predikcia priradená. Krátke úseky sa kontrolujú iba v prípade hélixov alebo listov, keďže u ohybu sa vždy do výsledného vektoru dosadzujú štvorice znakov. Výsledná predikcia pre zvolenú sekvenciu potom vyzerá nasledovne:

```
Strukt =
HHHHHHHHCCCCEEEEEEEEECEEEEECCCTTTTCTTTTTTTTEEEEEEEEEEEEETTTTCCCTTTTC
```

4.4 Grafické rozhranie

Grafické užívateľské rozhranie bolo vytvorené z časti programovo a z časti pomocou matlabovského prostredia GUIDE. Slúži predovšetkým ku grafickému zobrazeniu predikcie, záznamu predikovaných sekvencií a exportu výsledkov do súboru.

4.4.1 Popis komponentov

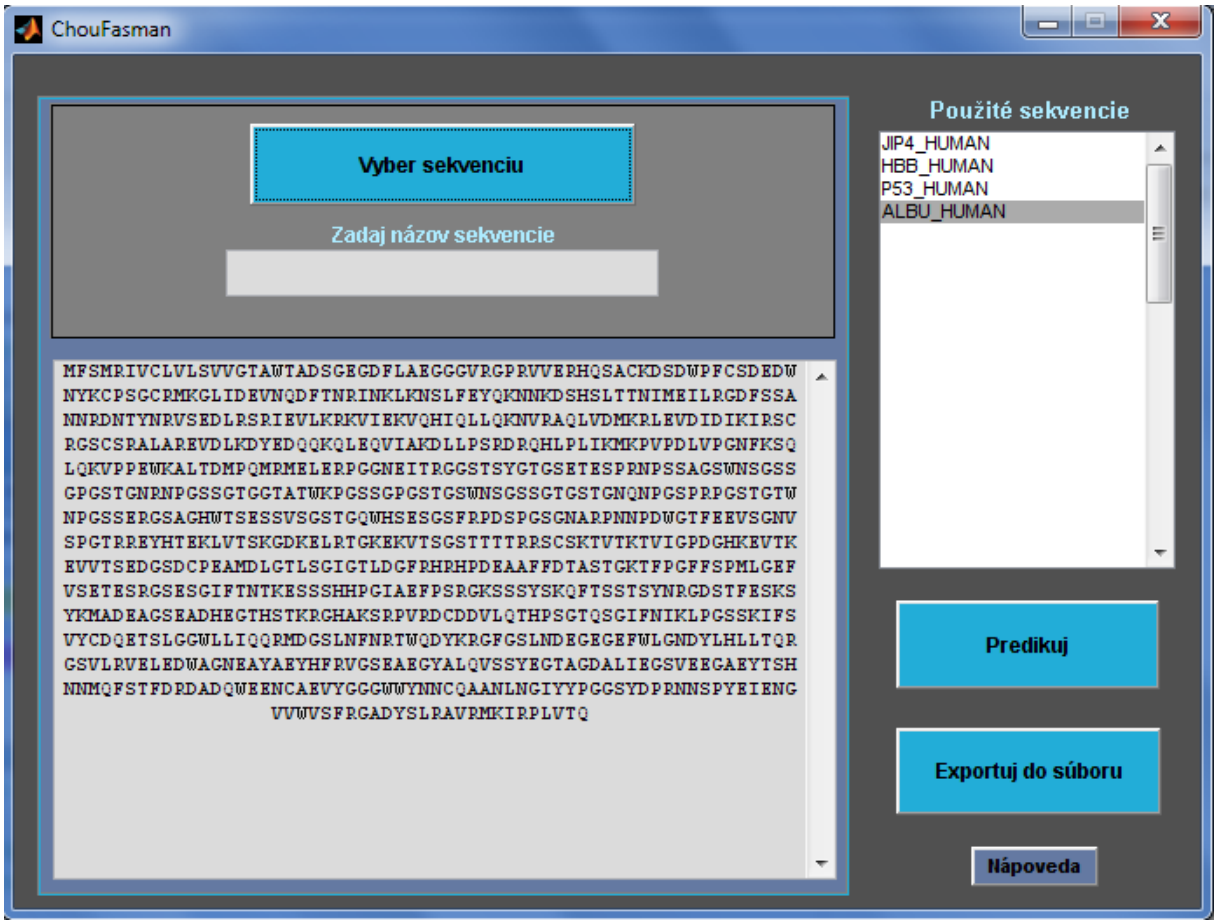


Obr. 14: Užívateľské prostredie programu ChouFasman

Tlačítko **Vyber sekvenciu** otvára dialógové okno, ktoré umožňuje vybrať a načítať požadovanú sekvenciu zo súboru. Aplikácia pracuje so súbormi výhradne vo formáte FASTA. Editovateľné pole nižšie slúži pre zobrazenie vybranej sekvencie alebo jej priame zadanie vo forme reťazca (postupnosti znakov) v štandardnom aminokyselinovom kódovaní.

Pole **Zadej názov sekvencie** priraďuje vybranej sekvencii názov, pod ktorým bude ďalej identifikovateľná. V prípade, že pole ostane nevyplnené, priradí sa sekvencii názov automaticky, a to nasledovne: pokiaľ nie je zadaný názov, ale je známy súbor, názov je získaný priamo z hlavičky

FASTA formátu. V opačnom prípade, pri zadaní do prázdneho poľa, je sekvencia označená *Bez názvu x*, x predstavuje poradové číslo predikcie.

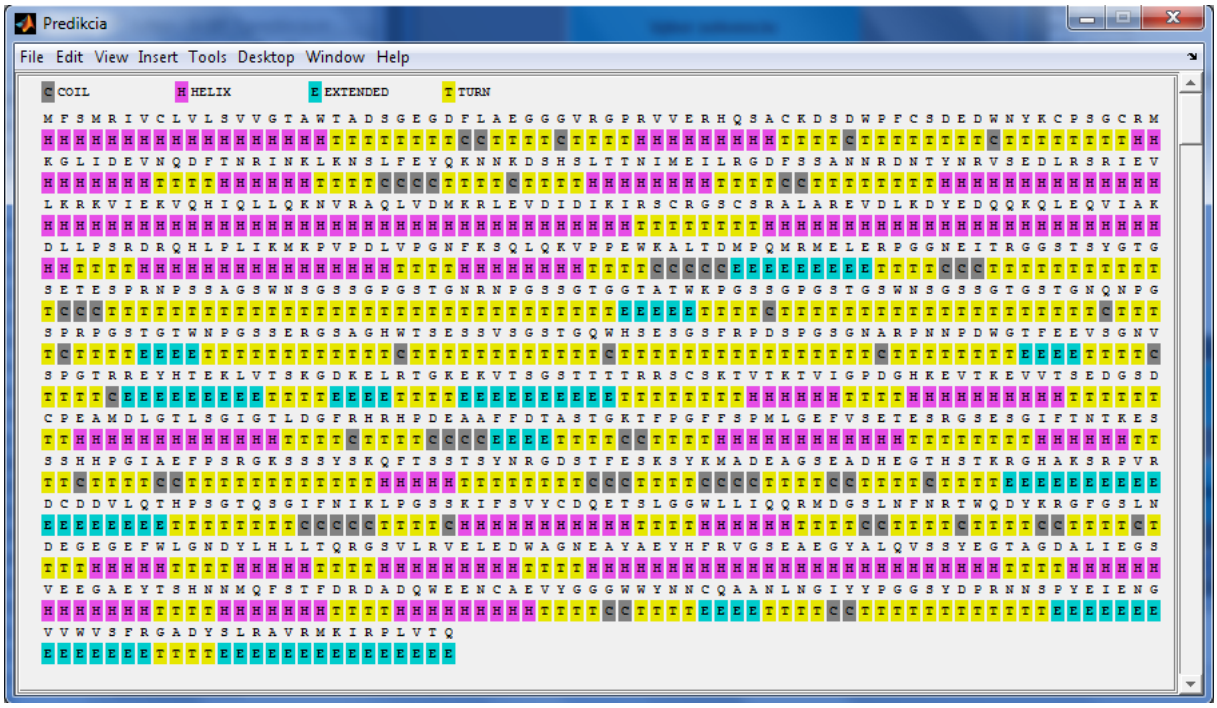


Obr. 15: Vzhľad hlavného okna po prevedení predikcie

Tlačítko **Predikuj** otvára samostatné okno s predikciou vytvorenou vyššie uvedenou funkciou `ChouFasman`. Tlačítko zároveň ukladá aktuálne spracovávanú sekvenciu do zoznamu, buď pod zadaným alebo automaticky priradeným názvom. Pre spustenie predikcie je nutné, aby bola zadaná aminokyselinová sekvencia.

Predikcia je zobrazená formou, kde každá časť pôvodnej sekvencie (70 reziduí) je nasledovaná odpovedajúcou sekundárnou štruktúrou s farebne vyznačenými oblasťami pre jednotlivé štruktúrne triedy podľa legendy (Obr. 16).

Zoznam **Použité sekvencie** obsahuje všetky sekvencie, pre ktoré bola predikovaná štruktúra od začiatku spustenia aplikácie. Výber položky zo zoznamu umožňuje zobrazenie sekvencie a jej názvu do príslušných okien a tiež opätovné vygenerovanie predikcie.

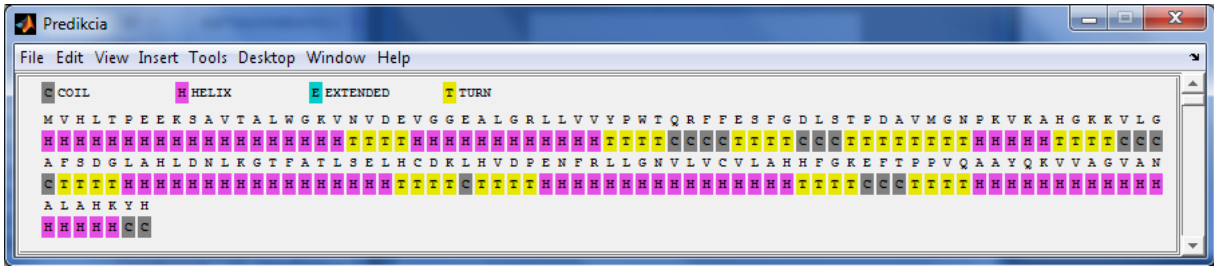


Obr. 16: Grafické zobrazenie predikcie programu v novom okne Predikcia

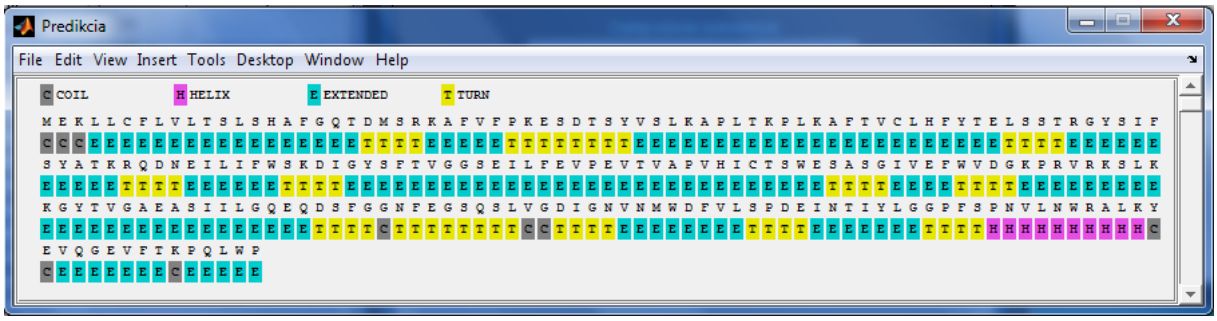
Tlačítko **Exportuj do súboru** otvára dialógové okno, ktoré umožňuje uloženie predikovaných dát do textového súboru. Súbor je defaultne ukladaný pod názvom príslušnej sekvencie a obsahuje názov proteínu, pôvodnú sekvenciu a predikovanú štruktúru. Pre vytvorenie súboru je potrebné, aby bola najprv predikovaná štruktúra požadovanej sekvencie.

Tlačítko **Nápoveda**, otvára samostantné okno, v ktorej je zobrazená nápoveda k používaniu aplikácie z textového súboru.

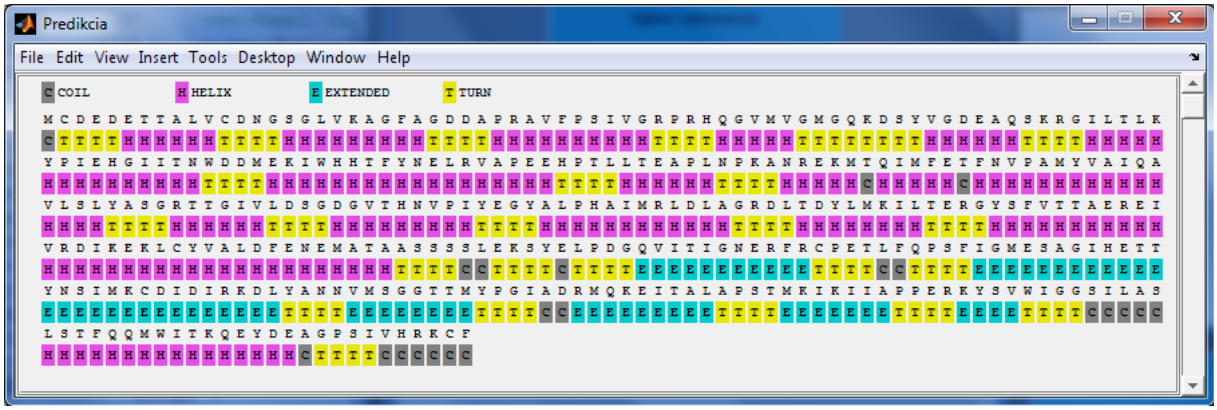
4.5 Výsledky



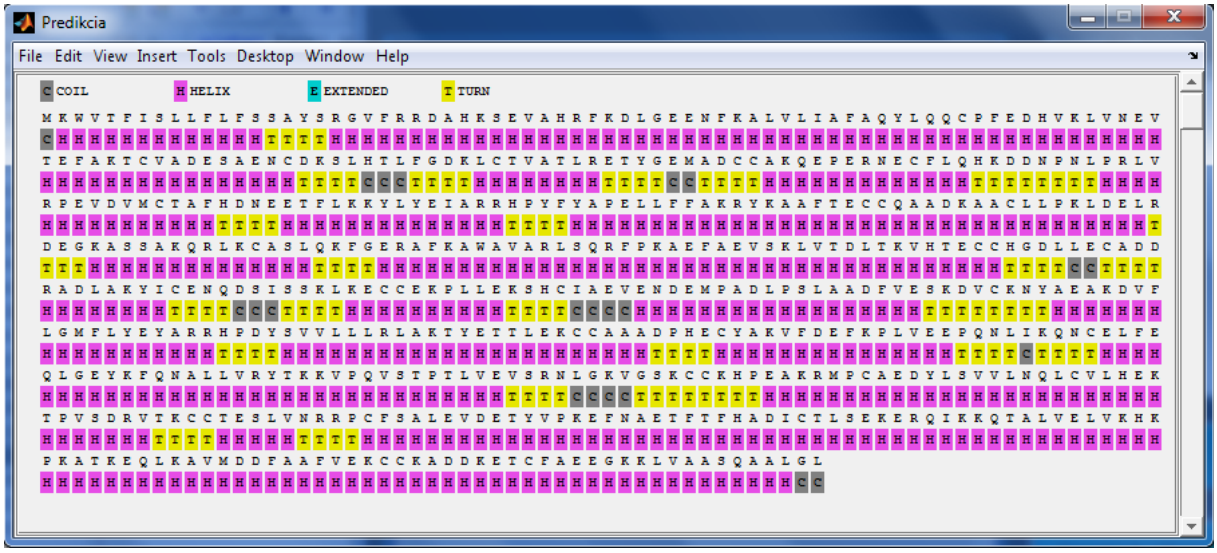
Obr. 17: Predikcia programu pre sekvenciu 1



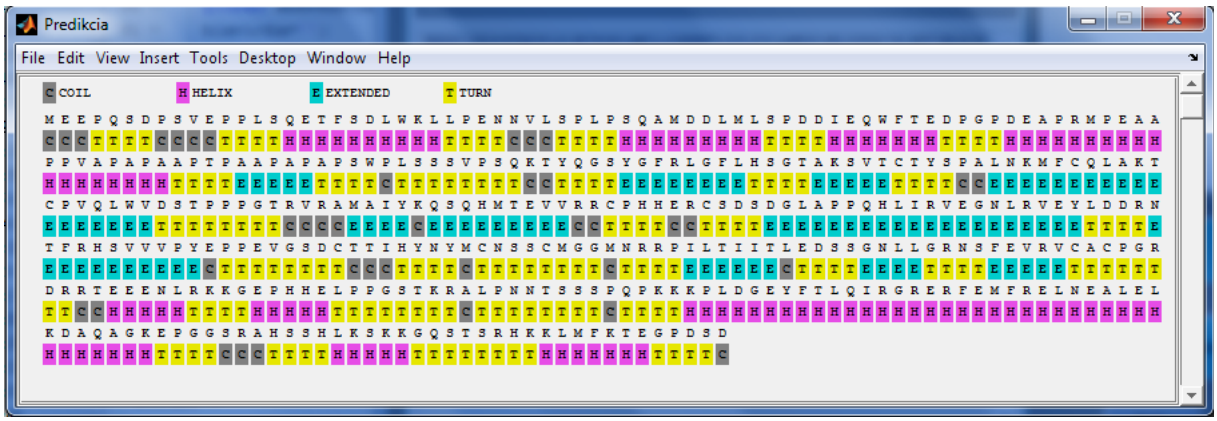
Obr. 18: Predikcia programu pre sekvenciu 2



Obr. 19: Predikcia programu pre sekvenciu 3



Obr. 20: Predikcia programu pre sekvenciu 4



Obr. 21: Predikcia programu pre sekveniu 5

4.6 Hodnototenie predikovaných výsledkov, výpočet presnosti predikcie

Pre objektívne zhodnotenie výsledkov predikčných metód a ich vzájomné porovnanie je nutné zaviesť parameter vyjadrujúci správnosť predikcie. Najčastejšie využívaným skóre je tzv. trojstavová presnosť Q_3, udávajúca percentáž všetkých správne predikovaných reziduií vzhľadom k známym štruktúrnym stavom z experimentov cez tri štruktúrne formy *t*:

$$Q_3 = \sum (t = H, E, C) \frac{predikované_t}{pozorované_t} \cdot 100 \quad , \tag{10}$$

kde *H* = hélix, *E* = skladaný list a *C* = náhodný zhluk/iná štruktúra.

Skóre je analogicky možné vypočítať i pre viac štruktúrnych tried, v praxi, pri hodnotení metód, je však Q_3 určitým štandardom. Napriek tomu, podáva predovšetkým prehľadovú informáciu o správnosti predikcie. Nie je úplne vhodná, pokiaľ cieľová štruktúrna trieda je prítomná iba v krátkom úseku sekvencie, vtedy jeho správna predikcia fiktívne zvyšuje celkové skóre. Sofistikovanejšou metódou je odhad Matthewho korelačného koeficientu podľa vzorca:

$$C = \frac{t_p t_n - f_p f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}} \quad , \tag{11}$$

kde t_p („true positive“) – počet správne pozitívnych predikcií, t_n („true negative“) – počet správne negatívnych predikcií, f_p („false positive“) – počet falošne pozitívnych predikcií a f_n („false negative“) – počet falošne negatívnych predikcií. Hodnota C sa pohybuje v intervale <0,1>.

Spomínané parametre hodnotia správnosť predikcie vzhľadom na samostatné reziduá. Pre zachytenie predikcie jednotlivých segmentov sa využíva SOV („Segment Overlap“) index, ktorý počíta s rozsahom, na ktorom sa štruktúry dvoch segmentov prekrývajú:

$$Sov = \frac{1}{N} \sum_{s} \frac{minov(s_1; s_2) + \delta}{maxov(s_1; s_2)} \cdot len(s_1) \quad , \tag{12}$$

kde N je celkový počet reziduí a s porovnávané segmenty, ktoré vykazujú pre aspoň jednu pozíciu rezidua rovnakú sekundárnu štruktúru (index 1 pre pozorovaný segment, index 2 pre predikovaný segment). *Minov* predstavuje počet reziduí, pre ktoré majú oba segmenty rovnakú štruktúru, tj. aktuálne prekrytie segmentov, *maxov* celý rozsah, ktorému je aspon u jedného z dvoch segmentov priradená daná štruktúra. Hodnota δ povoluje určité malé odchýlky pre nerovnaké reziduá na konci segmentov, ktoré sú bežné u štruktúrnych homológnych sekvencií. Váha $len(s_1)$ je dĺžka experimentálnej sekvencie a predstavuje asymetriu medzi oboma segmentami. [26], [27]

4.6.1 Hodnotenie metód

Kvalita predikcie, hodnotenie metód a ich vzájomné porovnávanie sa zakladá takmer výlučne na presnosti predikcie. Hodnoty parametru Q_3 pre jednotlivé metódy uvedené v práci sú zhrnuté v nasledujúcej tabuľke spolu s odkazom na literatúru, ktorá ich uvádza. Metódy sú na ich základe porovnateľné len orientačne, keďže každá z nich mala pravdepodobne inú množinu testovacích dát.

Tab. 3: Uvádzaná presnosť vybraných metód a presnosť vypočítaná pre realizovaný program

Metóda	Q_3 [%]	Hodnota prevzaná z
GOR I	55,00	[14]
GOR II	56,00	[14]
GOR III	63,30	[14]
GOR IV	64,40	[14]
základné PhD	70,80	[29]
fuzzy k-NN	75,75	[30]
ChouFasman	57,89	

Skutočná presnosť programov vo výsledku závisí na výbere databázy, spôsobe implementácie, nastavení parametrov metódy a zahrnutia ďalších vyhladzujúcich postupov, napr. rôznych druhov zarovnania sekvencií.

V tabuľke je zároveň uvedená vypočítaná hodnota presnosti realizovaného programu podľa metódy Chou-Fasman. Aj keď program umožňuje štvorstavovú predikciu, pre výpočet presnosti bola použitá redukcia do troch stavov ({T} → C), aby boli výsledky vzájomne porovnateľné. Presnosť bola získaná podľa vzťahu (10) na základe výsledkov predikcie (Tab. 4) pre celkový počet reziduí

N= 1750. Presnosť pre Chou-Fasmanov algoritmus uvádzaná v literatúre sa pohybuje v rozmedzí 50- 60%. Nedokonalosťou metódy je predovšetkým fakt, že nepočíta s vplyvom okolitých aminokyselín, predikcia prebieha pre samostatné aktuálne reziduum. Ďalším nedostatkom sú preferenčné parametre metódy, ktoré boli odvodené z relatívne malého súboru dát a empirické rozhodovacie pravidlá.

Tab. 4: Výsledky predikcie programu ChouFasman spolu s „pozorovanými“ hodnotami z databázy

	H		E		C	
	tp	**pozorované**	**tp**	**pozorované**	**tp**	**pozorované**
sekv. 1	73	112	0	0	22	35
sekv. 2	3	18	79	94	44	112
sekv. 3	71	122	7	49	79	206
sekv. 4	341	424	0	9	47	176
sekv. 5	40	58	49	83	154	252
Σ	528	734	135	235	346	781

Hodnoty *tp* („true positive“) v Tab. 4 predstavujú počet správne predikovaných reziduí pre konkrétnu štruktúru. Ako *pozorované* sú uvedené hodnoty na základe známych sekundárnych štruktúr z UNIPROT databázy. Výslednú presnosť však nie je možné brať reprezentačne, keďže nebol použitý štatisticky relevantný súbor dát.

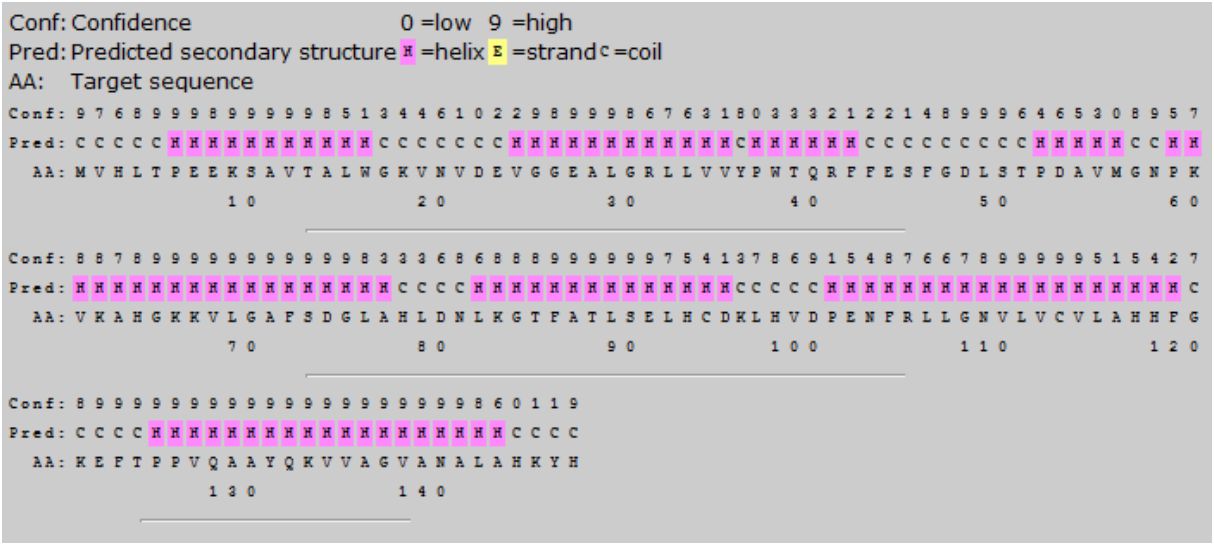
5 Prehľad vybraných softvérov

Nasledujúca kapitola je venovaná vybraným, voľne prístupným softvérom na internete a porovnaniu ich grafických výstupov a parametrov predikcie (čas, presnosť). Presnosť predikcie, udaná vo forme parametru Q_3 pre tri stavy (viď vyššie), je uvádzaná v dokumentáciách softvérov. Pre otestovanie bola použitá sekvencia 1 (HBB) a sekvencia 2 (CRP), najmä kvôli výrazne odlišnej sekundárnej štruktúre, ktorá pozostáva u oboch sekvenci takmer výlučne z jednej štruktúrnej formy. Predikované výsledky je teda možné orientačne zhodnotiť i vizuálne.

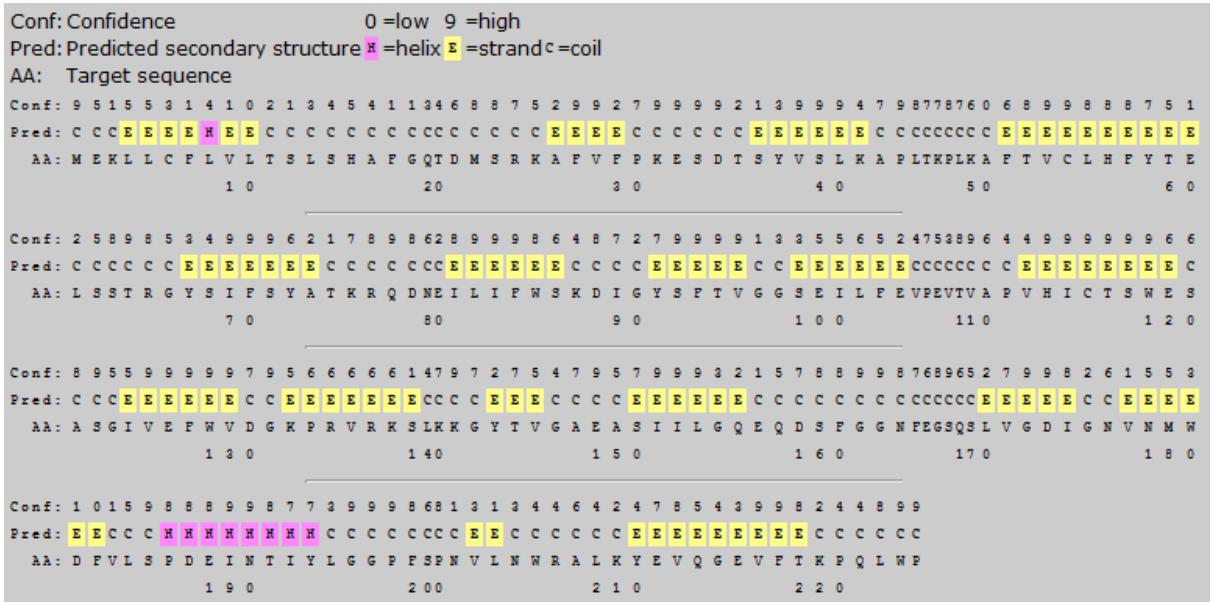
5.1 PSIPRED

- online predikcia a stiahnuteľný softvér
- metóda: dopredné neurónové siete, pre analýzu využívajú výstup získaný z PSI-BLAST zarovnania
- čas predikcie: cca 20 min
- uvádzaná presnosť: 80,7 % (Q_3)
- predikcia dostupná z: http://bioinf.cs.ucl.ac.uk/psipred/
- poznámka: navyše určuje i mieru istoty priradenia štruktúry

Výsledky:



Obr. 22: Predikcia programu PSIPRED pre sekvenciu 1

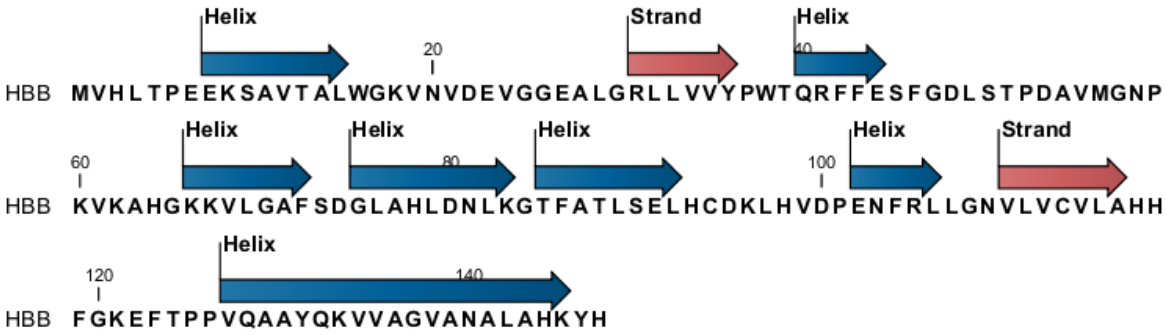


Obr. 23: Predikcia programu PSIPRED pre sekvenciu 2

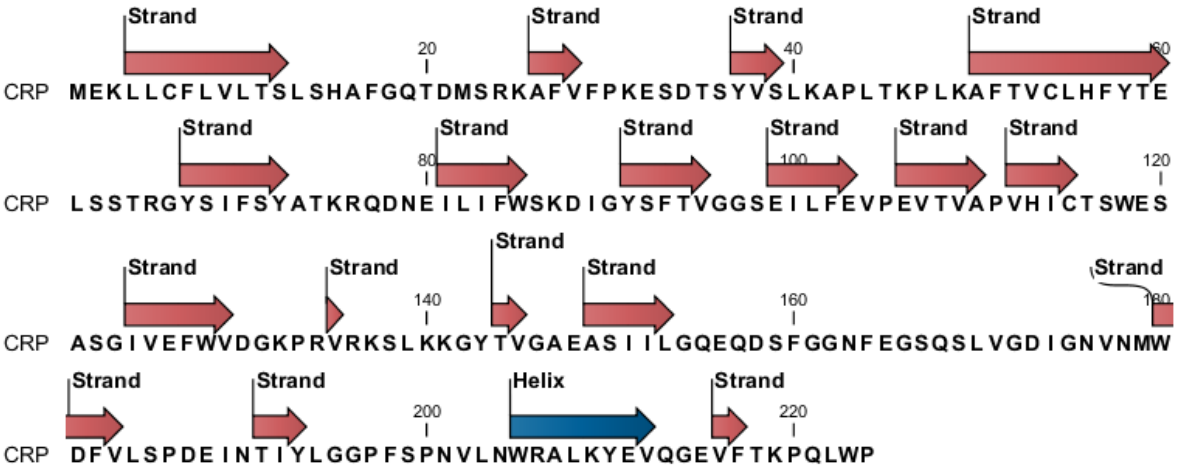
5.2 CLC Protein Workbench 5.6.1

- stiahnuteľný softvér
- metóda: Hidden Markov Model (HMM), modely trénované na dátach z PDB databázy
- čas predikcie: okamžite
- uvádzaná presnosť: 68% (Q3)
- softvér dostupný z: http://www.clcbio.com/index.php

Výsledky:



Obr. 24: Predikcia programu CLC Protein pre sekvenciu 1



Obr. 25: Predikcia programu CLC protein pre sekvenciu 2

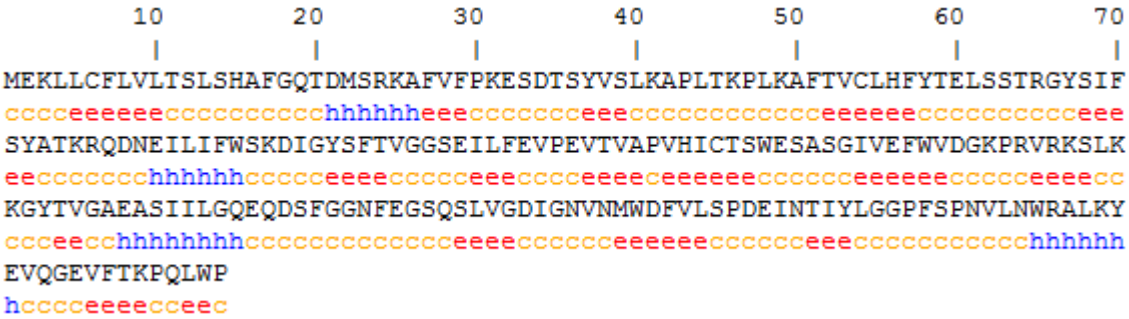
5.3 GOR IV

- online predikcia
- Metóda: oproti klasickej metóde GOR (viď str. 18) navyše uvažuje vplyv párových frekvencií výskytu reziduí v určitom konformačnom stave
- čas predikcie: cca 2-3 min
- uvádzaná presnosť: 64,4% (Q_3)
- predikcia dostupná z: http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html

Výsledky:

```
        10        20        30        40        50        60        70
         |         |         |         |         |         |         |
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
ccccccchhhhhhhhccceeeccccchhhheeeeeccccceeeeccccccccccccccccccccccceeee
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ccchhhhhhhccccchhhhhhhhccccccccchhhhhhceeeeeeeccccccccchhhhhhhhhhhhhhhh
ALAHKYH
hhcceec
```

Obr. 26: Predikcia programu GOR IV pre sekvenciu 1

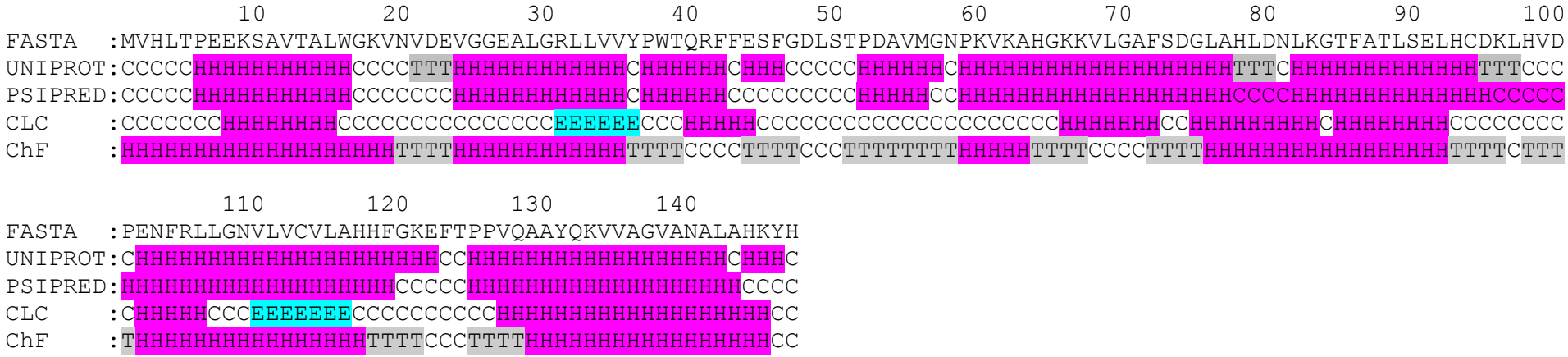


Obr. 27: Predikcia programu GOR IV pre sekvenciu 2

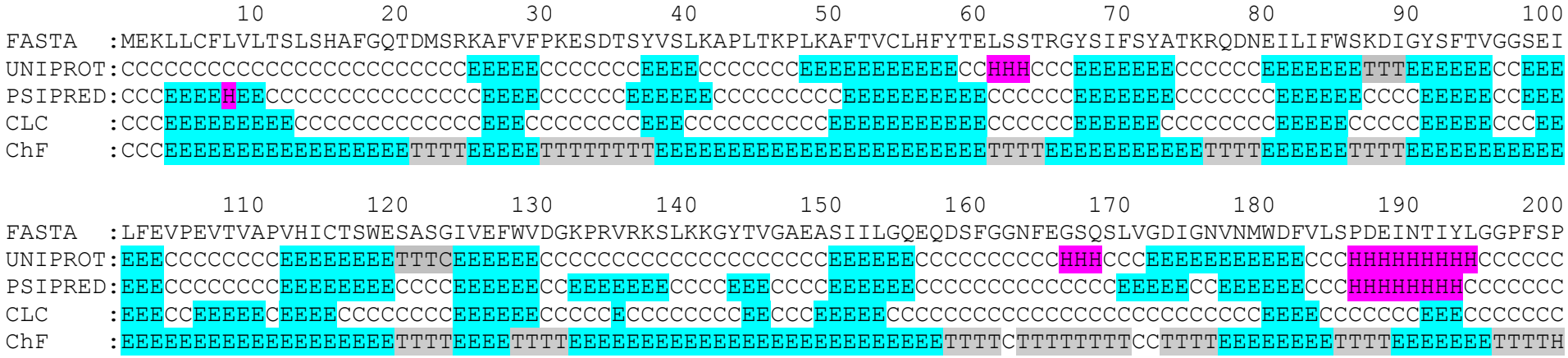
6 Záverečné grafické porovnanie výsledkov softvérov

H Helix E Skladaný list T Ohyb C Zhluk

Sekvencia 1:



Sekvencia 2:

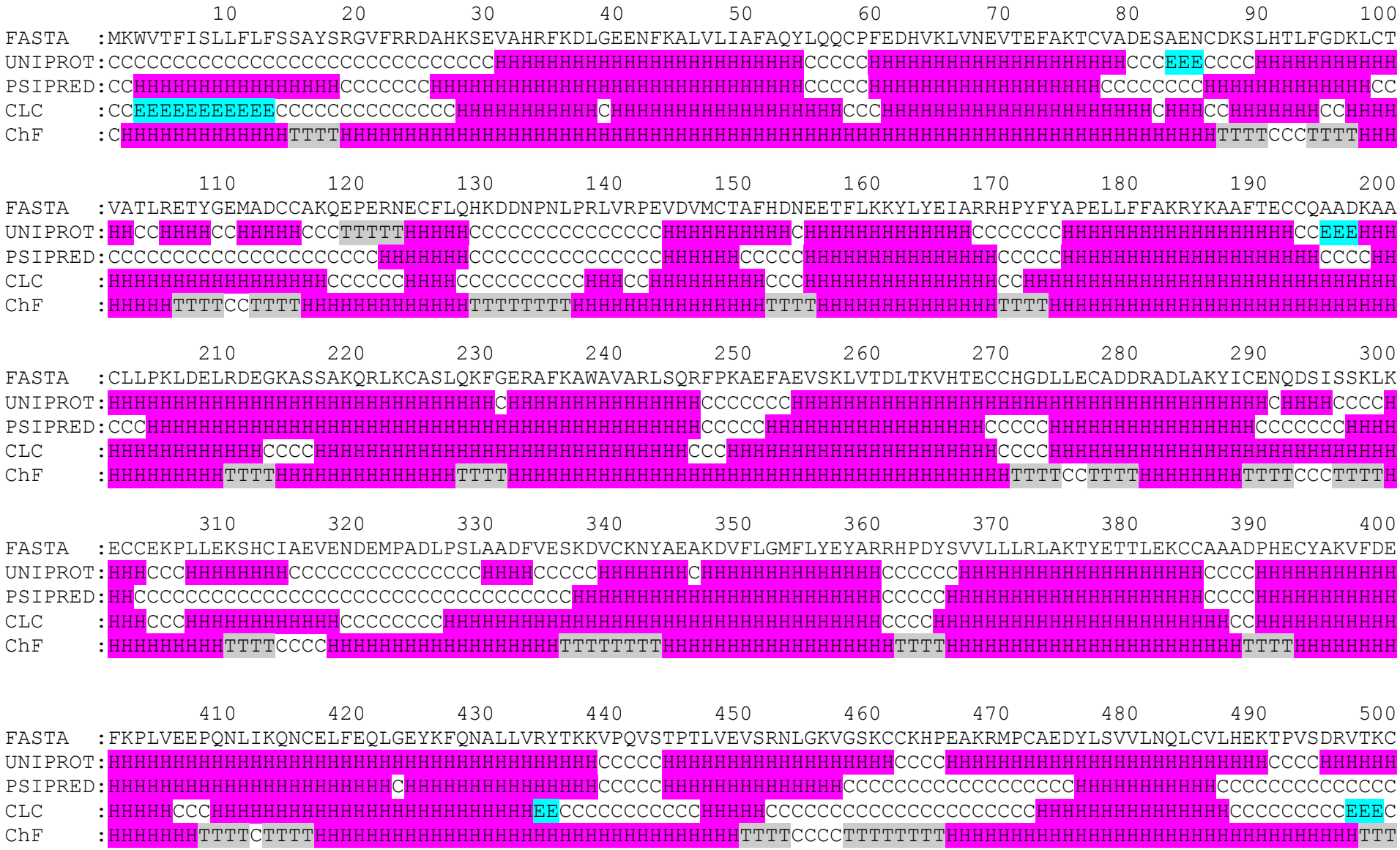


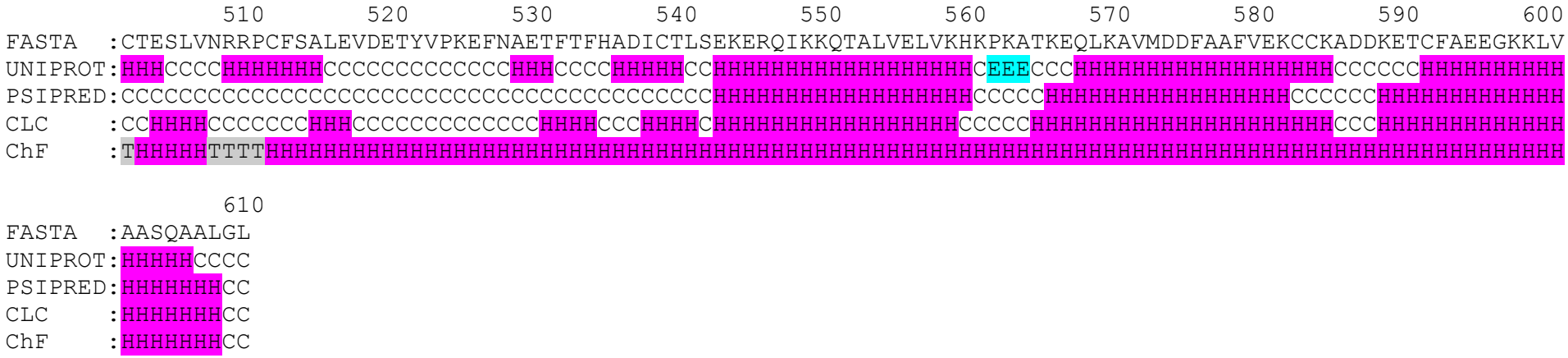
```
               210       220
FASTA   :NVLNWRALKYEVQGEVFTKPQLWP
UNIPROT:EEECHHHCCEEEEEECEEEEEEEE
PSIPRED:CEECCCCCCEEEEEEEEECCCCCC
CLC     :CCCCHHHHHHHHCCCEECCCCCCC
ChF     :HHHHHHHHHCCEEEEEEECEEEEE
```

Sekvencia 3:

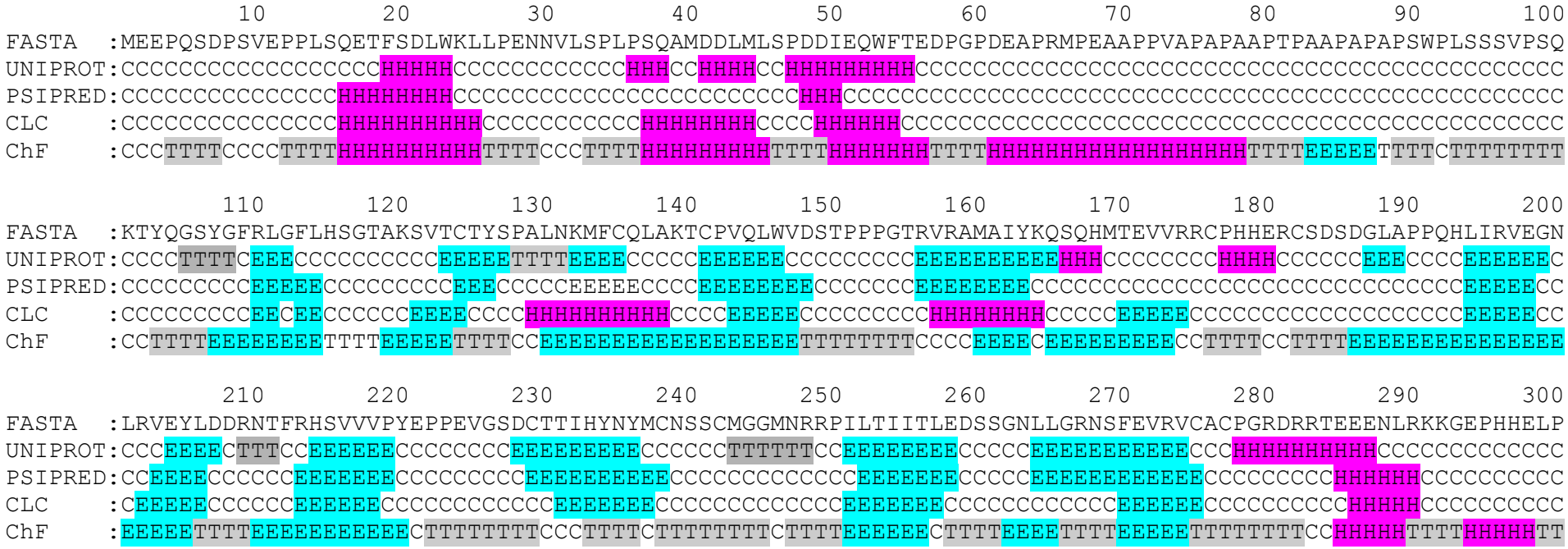
```
                10        20        30        40        50        60        70        80        90       100
FASTA   :MCDEDETTALVCDNGSGLVKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVGDEAQSKRGILTLKYPIEHGIITNWDDMEKIWHHTFYNELRVAP
UNIPROT:CCCCCCCCCEEEECCCCCEEECCCCCCCEEEECCCCCCCCCCCCCCCCCCCCCCCCCCTTTCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCCH
PSIPRED:CCCCCCCCEEEEECCCCCCEECCCCCCCCCEECCCCCCCCCCCCCCCCCCCCCCCCCHHHHHCCCCCEEECCCCCCCCCCHHHHHHHHHHHHCCCCCCCC
CLC     :CCCCCCCEEEEECCCCCCEECCCCCCCCCCCECCCECCCCCCCCEEECCCCCCCCCCCCCCCCCCEEEEECCCCCCEEECCCCHHHHHHHCECCCCECCC
ChF     :CTTTTHHHHHHTTTTHHHHHHHHHTTTTHHHHHHHHHHTTTTHHHHHTTTTTTTTHHHHHHTTTTHHHHHHHHHHHHHHHTTTTHHHHHHHHHHHHHHHH

               110       120       130       140       150       160       170       180       190       200
FASTA   :EEHPTLLTEAPLNPKANREKMTQIMFETFNVPAMYVAIQAVLSLYASGRTTGIVLDSGDGVTHNVPIYEGYALPHAIMRLDLAGRDLTDYLMKILTERGY
UNIPROT:HHCCEEEEECCCCCHHHHHHHHHHHHHCCCCEEEEEEEHHHHHHHHCCCCCCCEEEECCCCEEEECCCCCCCCHHHCEEECCCCCHHHHHHHHHHHHCCC
PSIPRED:CCCCCCCCCCCCCCHHHHHHHHHHHHCCCCCCEEEEEEHHHHHHHHCCCCEEEEEECCCCCEEEEECCCCCCCCCCEEEECCCHHHHHHHHHHHHHHCCC
CLC     :CCCCCEECCCCCCCCCCHHHHHHHHHHCCCCCCHHHHHHHHHHHHCCCCCCCEEECCCCCCECCCCEECCCCCCHHHHHHHHHHHHHHHHHHHHHHHHCC
ChF     :HHTTTTHHHHHHTTTTHHHHHCHHHHHCHHHHHHHHHHHHHHHHTTTTHHHHHHTTTTHHHHHHHHHTTTTHHHHHHHHHHHHTTTTHHHHHHHHTTTTH

               210       220       230       240       250       260       270       280       290       300
FASTA   :SFVTTAEREIVRDIKEKLCYVALDFENEMATAASSSLEKSYELPDGQVITIGNERFRCPETLFQPSFIGMESAGIHETTYNSIMKCDIDIRKDLYANNV
UNIPROT:CCCCHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHCCCCCEEEECCCCCEEEECCHHHHHHHHHHCHHHCCCCCCCCHHHHHHHHCCCCTTTHHHHHHCEE
PSIPRED:CCCCCHHHHHHHHHHHHCCCCCCCHHHHHHHHCCCCCCEEEEECCCCCEEEECCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHCCHHHHHHHHCCEE
CLC     :HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCEEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHCCCC
ChF     :HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHTTTTCCTTTTCTTTTEEEEEEEEEEETTTTCCTTTTEEEEEEEEEEEEEEEEEEEEEEEEEEETTTTE

               310       320       330       340       350       360       370
FASTA   :MSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWITKQEYDEAGPSIVHRKCF
UNIPROT:EECCCCCCCCCTTTTTTTTTCCCCCCCCCCCCCCTTTTCTTTCCCCHHHHCCCHHHHEEEHHHHHHCCCCHHHHHCC
PSIPRED:EECCCCCCCCCHHHHHHHHHHCCCCCCCEEEECCCCCCCCCCHHHHHHCCCCCCCCCCCCHHHHCCCCCCCEECCCC
CLC     :CCCCCCCCCCHHHHHHHHHHHHCCCCCEEEEECCCCCCEEEEECCCEHHHHHHHHHHHHHHHHCCCCCCCEECCCCC
ChF     :EEEEEEETTTTCCEEEEEEEEETTTTEEEEEEETTTTEEEETTTTCCCCCHHHHHHHHHHHHHHHHCTTTTCCCCCC
```

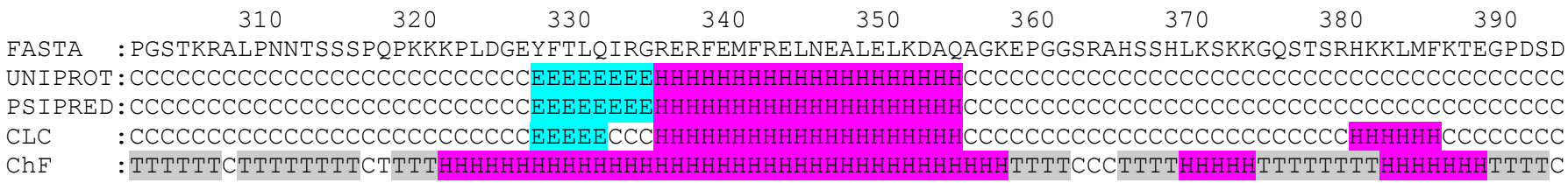
Sekvencia 4:


10 20 30 40 50 60 70 80 90 100
FASTA :MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENCDKSLHTLFGDKLCT
UNIPROT:CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHHCCCEEECCCCHHHHHHHHHH
PSIPRED:CCHHHHHHHHHHHHHHHCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHCC
CLC :CCEEEEEEEEEECCCCCCCCCCCCCCHHHHHHHHHHCHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHCHHHCCHHHHHHCCHHHH
ChF :CHHHHHHHHHHHHTTTTHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHTTTTCCCTTTTHHH
110 120 130 140 150 160 170 180 190 200
FASTA :VATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYLYEIARRHPYFYAPELLFFAKRYKAAFTECCQAADKAA
UNIPROT:HHCCHHHHCCHHHHHCCCTTTTTHHHHHCCCCCCCCCCCCCCCCHHHHHHHHHHCHHHHHHHHHHHHCCCCCCCHHHHHHHHHHHHHHHHHCCEEEHHH
PSIPRED:CCCCCCCCCCCCCCCCCCCCCCHHHHHHHCCCCCCCCCCCCCCCCHHHHHHCCCCCHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHCCCCHH
CLC :HHHHHHHHHHHHHHHHHCCCCCCHHHHCCCCCCCCCCHHHCCHHHHHHHHHCCCHHHHHHHHHHHHHHHHCCHHHHHHHHHHHHHHHHHHHHHHHHHH
ChF :HHHHHTTTTCCTTTTHHHHHHHHHHHHHTTTTTTTTHHHHHHHHHHHHHHHHTTTTHHHHHHHHHHHHHHTTTTHHHHHHHHHHHHHHHHHHHHHHHHH
210 220 230 240 250 260 270 280 290 300
FASTA :CLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKLVTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLK
UNIPROT:HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCHHHHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCHHHHCCCCH
PSIPRED:CCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHCCCCCCCCHHHH
CLC :HHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHHHHHHH
ChF :HHHHHHHHHTTTTHHHHHHHHHHHHHHHTTTTHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHTTTTCCTTTTHHHHHHHHTTTTCCCTTTTH
310 320 330 340 350 360 370 380 390 400
FASTA :ECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYAEAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKCCAAADPHECYAKVFDE
UNIPROT:HHHCCCHHHHHHHHCCCCCCCCCCCCCCCCHHHHCCCCCHHHHHHHCHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHH
PSIPRED:HHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHH
CLC :HHHCCCHHHHHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHHCCHHHHHHHHHH
ChF :HHHHHHHHHTTTTCCCCHHHHHHHHHHHHHHHHHTTTTTTTTHHHHHHHHHHHHHHHHHHTTTTHHHHHHHHHHHHHHHHHHHHHHTTTTHHHHHHHH
410 420 430 440 450 460 470 480 490 500
FASTA :FKPLVEEPQNLIKQNCELFEQLGEYKFQNALLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKC
UNIPROT:HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHHHHCCCCHHHHHH
PSIPRED:HHHHHHHHHHHHHHHHHHHHHHCHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCHHHHHHHHHHCCCCCCCCCCCCCC
CLC :HHHHHCCCHHHHHHHHHHHHHHHHHHHHHHHHEECCCCCCCCCCCHHHHHCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHCCCCCCCCCEEEC
ChF :HHHHHHHTTTTCTTTTHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHTTTTCCCCTTTTTTTTHHHHHHHHHHHHHHHHHHHHHHHHHHHHHTTT


510 520 530 540 550 560 570 580 590 600
FASTA :CTESLVNRRPCFSALEVDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLV
UNIPROT:HHHCCCCHHHHHHHCCCCCCCCCCCCCCHHHCCCCHHHHHCCHHHHHHHHHHHHHHHHHCEEECCCHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHH
PSIPRED:CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHH
CLC :CCHHHHCCCCCCCCHHHCCCCCCCCCCCCCCHHHHCCCHHHHCHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHH
ChF :THHHHHTTTTHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
610
FASTA :AASQAALGL
UNIPROT:HHHHHCCCC
PSIPRED:HHHHHHHCC
CLC :HHHHHHHCC
ChF :HHHHHHHCC

Sekvencia 5:


10 20 30 40 50 60 70 80 90 100
FASTA :MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQ
UNIPROT:CCCCCCCCCCCCCCCCCCHHHHHCCCCCCCCCCCCCHHHCCHHHHCCHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
PSIPRED:CCCCCCCCCCCCCCCCHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CLC :CCCCCCCCCCCCCCCCHHHHHHHHHHCCCCCCCCCCCCHHHHHHHHCCCCHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
ChF :CCCTTTTCCCCTTTTHHHHHHHHHHTTTTCCCTTTTHHHHHHHHHTTTTHHHHHHHTTTTHHHHHHHHHHHHHHHHHTTTTEEEEETTTTCTTTTTTTT
110 120 130 140 150 160 170 180 190 200
FASTA :KTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGN
UNIPROT:CCCCTTTTCEEECCCCCCCCCCEEEEETTTTEEEECCCCCEEEEEECCCCCCCCCEEEEEEEEEHHHCCCCCCCCHHHHCCCCCCEEECCCCEEEEEEC
PSIPRED:CCCCCCCCCEEEEECCCCCCCCCEEECCCCCEEEEECCCCEEEEEEEECCCCCCCEEEEEEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEECC
CLC :CCCCCCCCCEECEECCCCCCEEEECCCCHHHHHHHHHHCCCCEEEEECCCCCCCCCCHHHHHHHHCCCCCEEEEECCCCCCCCCCCCCCCCCCCCEEEEECC
ChF :CCTTTTEEEEEEEETTTTEEEEETTTTCCEEEEEEEEEEEEEEEEETTTTTTTTCCCCEEEECEEEEEEEECCTTTTCCTTTTEEEEEEEEEEEEEEE
210 220 230 240 250 260 270 280 290 300
FASTA :LRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELP
UNIPROT:CCCEEEECTTTCCEEEEEECCCCCCCCEEEEEEEECCCCCCTTTTTTCCEEEEEEECCCCCEEEEEEEEECCCHHHHHHHHHCCCCCCCCCCCCC
PSIPRED:CCEEEECCCCCCEEEEEEECCCCCCCCCEEEEEEEEECCCCCCCCCCCCCCEEEEEEECCCCCEEEEEEEEEECCCCCCCCCHHHHHHCCCCCCCCCC
CLC :CEEEEECCCCCCEEEEEECCCCCCCCCCCCEEEEEEECCCCCCCCCCCCCCEEEEEEECCCCCCCCCCCCEEEEEECCCCCCCCCCHHHHHCCCCCCCCCC
ChF :EEEEETTTTEEEEEEEEEECTTTTTTTTCCCTTTTCTTTTTTTTCTTTTEEEEEECTTTTEEEETTTTEEEEETTTTTTTTCCHHHHHTTTTHHHHHTT


310 320 330 340 350 360 370 380 390
FASTA :PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
UNIPROT:CCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEEEHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
PSIPRED:CCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEEEHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CLC :CCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEECCCHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHCCCCCCCCC
ChF :TTTTTTCTTTTTTTTCTTTTHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHTTTTCCCTTTTHHHHHTTTTTTTTHHHHHHHTTTTC

6.1 Dodatok k záverečnému grafickému porovnaniu

Porovnanie a správnosť predikovaných výsledkov sa vzťahuje k dátam z databázy so známymi štuktúrami, tj. k riadku označnému UNIPROT. Keďže testované softvéry z internetu umožňujú iba trojstavovú predikciu, pre objektívne porovnanie musí byť prevedená redukcia týchto dát ({T} → C). V záverečnou zhrnutí boli použité dva z uvedených softvérov, ktoré mali vyššiu uvádzanú presnosť – PSIPRED a CLC BIO. Podľa teoretických predpokladov, resp. podľa hodnoty presnosti predikcie, by mal najlepšie výsledky dosahovať softvér PSIPRED. Tento predpoklad sa podarilo overiť i v praxi. Na použitých testovaných dátach správne predikoval 75,75% všetkých hélixov, 79,57% listov a po prevedení redukcie 83,10% zhlukov. Softvér CLC BIO odhadol správne 72,34% hélixov, 48,09% listov a 72,86% zhlukov.

Celková vypočítaná presnosť realizovaného pogramu ChouFasman a správne predikované počty sú uvedené v Tab. 3 a Tab. 4. Pre prehľadnosť možno uviesť percentuálne vyjadrenie „true positive“ predikcií: 71,93% hélixov, 57,45% listov a 44,30% zhlukov. Problematickou časťou realizovanej metódy je nadmerná predikcia ohybov neúmerná ich skutočnému výskytu. Väčšina softvérov však výskyt ohybov vôbec neuvažuje, sú zahrnuté v rámci triedy „náhodný zhluk“ – nemajú teda vplyv na udávanú presnosť Q_3.

Záver

V práci sú spracované teoretické náležitosti týkajúce sa proteínov vrátane zloženia, chemických vlastností a väzieb na molekulárnej úrovni, ktoré majú vplyv na usporiadanie do výslednej štruktúry proteínu. Jednotlivé štruktúry sú spracované s dôrazom na sekundárnu (2D) štruktúru, ktorej sa následne týkajú predikčné algoritmy.

Metodika skúmania proteínových štruktúr je rozdelená do dvoch častí. Podkapitola Experimentálne metódy popisuje základný princíp vytvárania štruktúrnych modelov na zakláde usporiadania atómov v makromolekule. Toto usporiadanie sa zisťuje prevažne pomocou RTG žiarenia a nukleárnej magnetickej rezonancie. Podkapitola Výpočtové metódy sa venuje štyrom metódam, ktoré predstavujú štyri hlavné prístupy k predikcii sekundárnej štruktúry. Súčasný trend smeruje k využívaní umelej inteligencie a rôznym zdokonaleniam (napr. viacnásobnému zarovnaniu sekvencií), prípadne k hybridizácii základných metód.

Praktická časť práce sa venuje programovej realizácii Chou-Fasmanovej metódy, ktorá je testovaná na súbore piatich ľudských proteínov z databázy UNIPROT a doplnená o jednoduché užívateľské rozhranie. Na rovnakom súbore dát boli zároveň testované tri voľne dostupné softvéry z internetu založené na odlišných metódach predikcie. Výsledky týchto softvérov sú vzájomne graficky porovnané s vlastným programom v záverečnej časti práce. Na základe výsledkov porovnania sa potvrdil teoretický predpoklad o miere presnosti jednotlivých softvérov. Celková presnosť Q_3 pre realizovaný program a daný súbor dát vyšla 57,89 %, ako najdokonalejšia sa ukázala predikcia α- hélixov (71,93 %). Metóda a zároveň program umožňujú v porovnaní s testovanými softvérmi navyše predikciu ohybov, ktorá však nie je príliš presná. Ohyby však predstavujú „doplňujúce“ sekundárne štruktúry, takže odchýlky sú viac akceptovateľné, ako u predikcie hlavných štruktúr.

Samotnú metódu predikcie by pre dosiahnutie väčšej presnosti bolo vhodné vylepšiť zahrnutím vplyvu lokálnych interakcií, čo už v podstate realizuje metóda GOR. Zároveň by bolo možné vytvoriť pozičné preferencie nielen pre ohyb, ale aj pre α-hélix a β-skladaný list, keďže niektoré reziduá majú tendenciu vyskytovať sa v určitých štruktúrach na špecifických pozíciach v reťazci. Príkladom je prolín, ktorý zvykne vytvárať tzv. „N-cap“. V prvých troch pozíciach od N- konca sa toto reziduum podieľa na štruktúre hélixu, aj keď v iných častiach reťazca funguje ako „helix breaker“.

Literatúra

[1] LESK, Artur M. *Introduction to Protein Architecture*: *The Structural Biology of Proteins*. NewYork: Oxford University Press, 2001, 360 s. ISBN 9780198504740

[2] DOSTÁL, Jiří , Hana PAULOVÁ, Jiří SLANINA a Eva TÁBORSKÁ. *Biochemie pro bakaláře*. Brno: Masarykova univerzita, 2003, 174 s. ISBN 80-210-3232-4.

[3] CREIGHTON, Thomas E. *Proteins*: *Structures and Molecular Properties*. 2. vyd. New York: W.H. Freeman, 1993, 512 s. ISBN 0-7167-1566-X.

[4] *Amino Acids* [online]. 2011 [cit. 2011-12-30]. Obrázok dostupný z: http://aminoacids.tk/new/bodybuilding-and-amino-acids/structure-of-amino-acid

[5] NEČAS, Oldřich. *Obecná biologie pro lékařské fakulty*. 3. vyd. Jinočany: H & H, 2000, 554 s.ISBN 80-86022-46-3.

[6] PETSKO, Gregory A. a Dagmar RINGE. *Protein Structure and Function*. London: New Science Press, 2004. ISBN 1-4051-1922-5.

[7] VOET, Donald J. a Judith G. VOET. *Biochemistry*. 3. vyd. John Wiley & Sons, 2004, 1616 s.ISBN 0-471-19350-X.

[8] BRANDEN, Carl a John TOOZE. *Introduction to Protein Structure*. 2. vyd. New York: Garland Publishing, 1998. ISBN 0-8153-2304-2.

[9] DERIS, Safaai Bin, Rosli Bin Md ILLIAS, Sahidan Bin SENAFI, Saad Osman ABDALLA a Satya Nanda Vel ARJUNAN. Protein Secondary Structure Prediction from Amino Acids Sequence Using Artificial Intelligence Technique. [online]. [cit. 2011-12-30]. VOT: 74017. Dostupné z: http://eprints.utm.my/4265/1/74017.pdf

[10] SINGH, Manpreet, Parvinder Singh SANDHU a Reet Kamal KAUR. Protein Secondary Structure Prediction. *World Academy of Science, Engineering and Technology* [online]. 2008 [cit. 2011-12-30]. Dostupné z: http://www.waset.org/journals/waset/v42/v42-87.pdf

[11] KRYSTEK, Stanley R., William J. METZLER a Jiri NOVOTNY Protein Secondary Structure Prediction: Computational Analysis. *Current Protocols in Protein Science*. John Wiley & Sons, 2000, 2.3.1-2.3.20. eISBN 9780471140863.
Dostupné z: http://www.nshtvn.org/ebook/molbio/Current%20Protocols/CPPS/ps0203.pdf

[12] SINGH, Rajbir, DEOL a Parvinder S. SANDHU. Chou-Fasman method for Protein Structure Prediction using Cluster Analysis. *World Academy of Science and Technology*. 2010, s.982-987.
Dostupné z: http://www.waset.org/journals/waset/v72/v72-178.pdf

[13] MOUNT, David W. *Bioinformatics*: *Sequence and Genome Analysis*. 2. vyd. New York: Cold Spring Harbor Laboratory Press, 2004, 692 s. ISBN 0-87969-687-7.

[14] Secondary Structure Considerations: GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence. ABELSON, John N. *Methods In Enzymology, Vol. 266*. New York: Academic Press, 1996, s. 540-553. ISBN 0121821676.
Dostupné z: http://www.biosyn.com/Images/ArticleImages/pdf/GOR.pdf

[15] KUMAR, Binod a N.N. JANI. Prediction of Protein Secondary Structure based on GOR Algorithm Integrating with Multiple Sequence Alignment. *International Journal of Advanced Engineering*. 2010(1). Dostupné z: http://steps-india.com/ijaea/18.pdf

[16] WÜTHRICH, Kurt. NMR Studies of Structure and Function of Biological Macromolecules. *Nobel Lecture* . 2002(12), 235-267. Dostupné z: http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2002/wutrich-lecture.pdf

[17] KRATOCHVÍL, Bohumil, Michal HUŠÁK, Jiří BRYNDA a Juraj SEDLÁČEK. Co nabízí současná RTG strukturní analýza?.*Chemické listy* [online]. 2008, Vol. 120, 889-901 [cit. 2011-12-30]. Dostupné z: http://www.chemicke-listy.cz/docs/full/2008_10_889-901.pdf

[18] WISHART, David . Lecture 3.1: Methods in 3D Structure Determination. *Proteomics* [online]. [cit. 2011-12-30]. Dostupné z: http://www.docstoc.com/docs/26257173/Protein-Structure Determination-by-X-ray-Crystallography

[19] BERG, Jeremy M., John L. TYMOCZKO a Lubert STRYER. *Biochemistry*. 6. vyd. New York: W. H. Freeman, 2007, 1025 s. ISBN 0-7167-8724-5.

[20] WHITFORD, David . *Proteins* : *Structure and Function*. Chichester: John Wiley & Sons, 2005, 528 s. ISBN 0-471-49893-9.

[21] Chou-Fasman Prediction of the Secondary Structure of Proteins: The Chou-Fasman-Prevelige Algorithm. PREVELIGE, Peter, Jr. a Gerald D. FASMAN *Prediction of Protein Structure and the Principles of Protein Conformation*. New York: Plenum Press, 1989, s. 392-416. ISBN 978-0-306-43131-9. Dostupné z: http://mail.informatika.org/~henny/Bioinformatics/ClassiPapers/preveligeandfasman1989.pdf

[22] Bioinformatics. *Imb-jena* [online]. [cit. 2011-12-30]. Obrázok dostupný z: http://www.imb-jena.de/~rake/Bioinformatics_WEB/basics_peptide_bond.html

[23] Peptide Bonds and Protein Structure. *Washington University in St. Louis* [online]. 2011 [cit.2012-05-22]. Obrázok dostupný z: http://www.nslc.wustl.edu/courses/Bio2960/labs/02Protein_Structure/PS2011.htm

[24] HORTON, Robert A. et al *Principles of Biochemistry 4/E*. New Jersey: Prentice Hall, 2005. ISBN 0131453068. Obrázok dostupný z: http://sandwalk.blogspot.com/2008/03/levels-of-protein-structure.html

[25] Photon Science. *Hasylab.desi* [online]. 2007/01/21 [cit. 2011-12-30]. Obrázok dostupný z: http://hasylab.desy.de/user_info/available_instruments/x_ray_protein_crystallography/index_eng.html

[26] ROST, Burkhard a Chris SANDER. Prediction of Protein Secondary Structure at Better than 70% Accuracy. *Journal of Molecular Biology*. 1993, Vol. 232, No. 2, s. 584-599. ISSN 0022-2836.

[27] DE HAAN a Jack A.M. LEUNISSEN *Protein Secondary Structure Prediction Comparison of Ten Common Prediction Algorithms Using a Neural Network*. Lansdale: IOS Press, 2005, s. 149-161. ISBN 978-1-58603-539-6. Dostupné z: http://www.bioinformatics.nl/~jackl/deHaan_EIB_2005.pdf

[28] ROST, Burkhard a Chris SANDER. Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure. *PROTEIN: Structure, Function and Genetics*. 1994, č. 19, s. 55-72.

[29] YI, Tau-Mu a Eric S. LANDER. Protein Secondary Structure Prediction Using Nearest-neighbor Methods. *Journal of Molecular Biology* [online]. 20. August 1993, Vol. 232, Issue 4, s. 1117-1129 [cit. 2012-05-22]. ISSN 0022-2836.
Dostupné z: http://www.sciencedirect.com/science/article/pii/S0022283683714646

[30] BONDUGULA, RAJKUMAR, OGNEN DUZLEVSKI a DONG XU. Profiles and Fuzzy k-Nearest Neighbor Algorithm for Protein Secondary Structure Prediction [online]. [cit. 2012-05-22]. Dostupné z: https://iiwas.comp.nus.edu.sg/~wongls/psZ/apbc2005/camera-ready/216.pdf

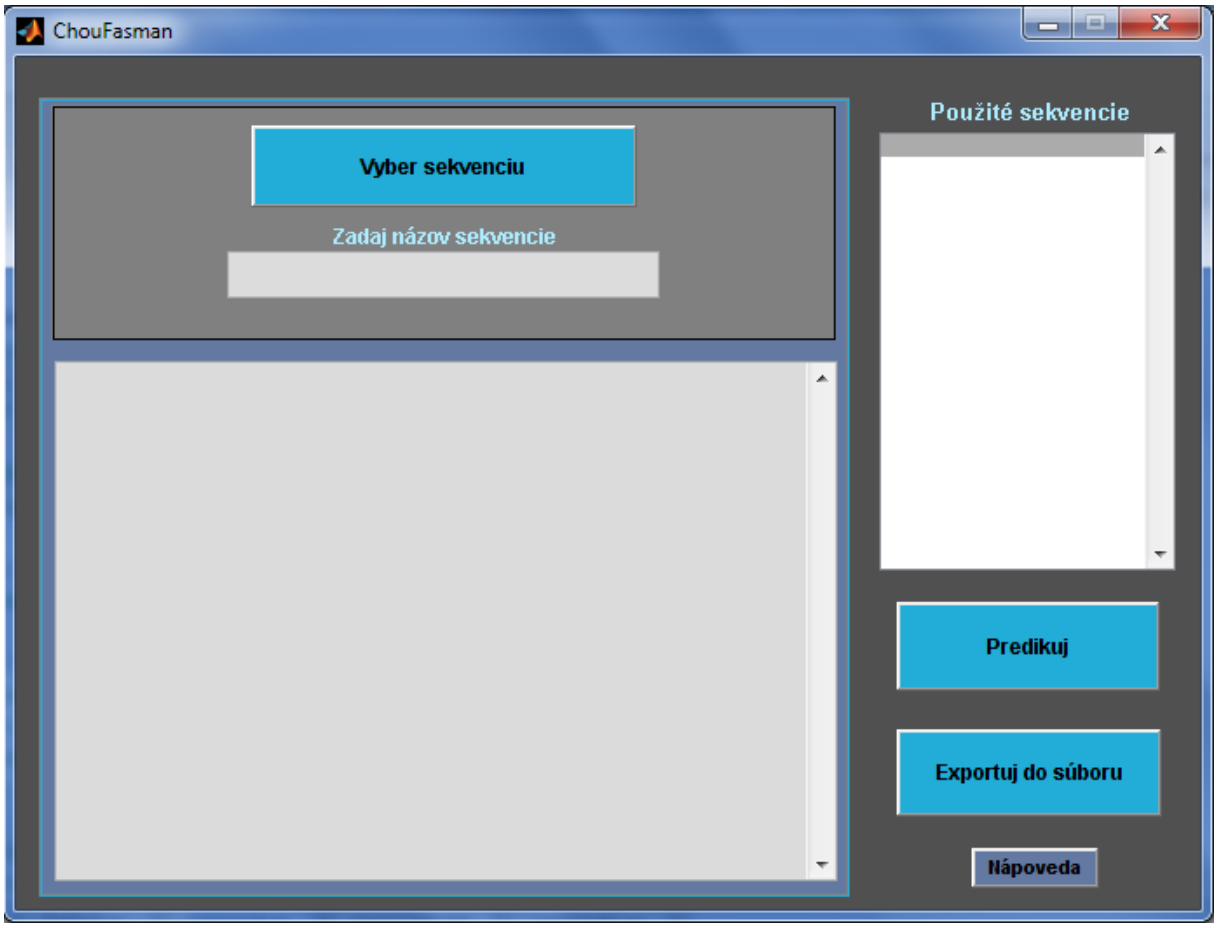
[31] LEDDA, Filippo Giuseppe. *Protein Secondary Structure Prediction: Novel Methods and Software Architectures*. Cagliari, 2011. Ph.D. Thesis. Dept. of Electrical and Electronic Engineering University of Cagliari. Supervisor Prof. Giuliano Armano. Dostupné z: http://veprints.unica.it/643/

[32] HERINGA. Computational Methods for Protein Secondary Structure Prediction Using Multiple Sequence Alignments.*Current Protein and Peptide Science*. 2000, Vol. 1, No. 3, s. 273-301. Dostupné z: http://www.ibi.vu.nl/teaching/masters/prot_struc/2005/heringa_cpps_1_273_301.pdf

Zoznam skratiek a symbolov

AA	aminokyselina
C	uhlík
CASP	Critical Assessment of Techniques for Protein Structure Prediction
CCD	Charged-Couple Device
COSY	Correlation Spectroscopy
DNA	deoxyribonukleová kyselina
DSSP	Database of Protein Structure Assignment
GOR	Garnier-Osguthorpe-Robson
H	vodík
MSA	Multiple Sequence Alignment
N	dusík
NMR	nukleárna magnetická rezonancia
NN	Nearest Neighbor
NOE	Nuclear Overhaus Effect
O	kyslík
PDB	Protein Data Bank
PSSP	Position Specific Scoring Matrix
RNA	ribonukleová kyselina
RTG	rentgen
SOV	Segment Overlap Score

Príloha – Užívateľský manuál



Program sa spúšťa zo súboru predikcia.m stlačením F5. Zobrazí sa hlavné okno aplikácie (viď. obrázok). Požadovanú sekvenciu určenú na predikcie vložte jedným z dvoch spôsobov:

1. stlačením tlačítka **Vyber sekvenciu** – otvorí sa dialógové okno pre výber súboru sekvencie vo formáte FASTA. Samotná sekvencia sa zobrazí do editovacieho poľa nižšie
2. priamym zadaním sekvencie vo forme textového reťaza do editovacieho poľa

Vyplnenie poľa **Zadaj názov sekvencie** je voliteľné. Priraďuje sekvencii názov, pod ktorým bude ďalej vystupovať. V prípade, že ostane nevyplnené, sekvencia bude označená automaticky podľa spôsobu zadania sekvencie:

1. v prípade načítania sekvencie zo súboru je označenie získané priamo z hlavičky FASTA formátu
2. pri priamom zadaní bude sekvencia označená *Bez názvu x*, x je poradovým číslom predikovanej sekvencie

Predikcia načítanej sekvencie sa spúšťa pomocou tlačítka **Predikuj**. V novom okne sa zobrazí predikcia podľa uvedenej legendy. Pôvodná sekvencia je rozdelená na časti a zobrazená v riadkoch bez farebného vyznačenia. Za každou časťou nasleduje príslušná predikcia zodpovedajúca jednotlivým aminokyselinám sekvencie. Stlačenie tlačítka zároveň uloží sekvenciu, s ktorou sa pracovalo, spolu s názom do listboxu **Použité sekvencie**.

Po predikovaní je možné predikciu spolu s názvom a pôvodnou sekvenciou uložiť do textového súboru. Stlačením tlačítka **Exportuj do súboru** sa otvorí dialógové okno pre uloženie vytvoreného súboru, defaultne pod názvom priradeným sekvencii. Exportuje sa vždy aktuálna položka zo zoznamu použitých sekvencií

Po stlačení lubovoľnej položky zoznamu **Použité sekvencie** je možné opätovne vygenerovať jej predikciu a zobraziť pôvodnú sekvenciu s názvom, prípadne dodatočne exportovať do súboru.