

Kapitola 2

Proteíny a predikcia ich štruktúry

Jacob Berzelius, významný švédsky chemik, v roku 1838 navrhol názov „proteín“ pre zložené organické zlúčeniny bohaté na dusík, ktoré nachádzal v bunkách živočíchov a rastlín. Štruktúra a chémia proteínov sa vyvinula a doladila behom milárd rokov evolúcie. Z chemického hľadiska sú proteíny najzložitejšie a funkčne najdomyselnejšie známe molekuly. Sú tvorené sekvenciou aminokyselín, ktorých väčšina bola objavená v priebehu 19. storočia. Sú pospájané kovalentnou peptidovou väzbou o veľkosti približne 1,32 Å, ktorá viaže väčšinu suchej hmotnosti bunky [3]. Vysoká početnosť funkcií, ktoré proteíny zaisťujú, pramení z obrovského počtu rôznych tvarov, ktorých môžu v priestore nadobúdať – funkcia sleduje štruktúru.

2.1 Štruktúra proteínov a genetický kód

Zásadným krokom v štúdiu proteínov bola práca Jamesa B. Sumnera z roku 1926, ktorý ukázal, že enzýmy možno kryštalizovať a izolovať. V roku 1953 Francis Crick a James D. Watson popísali štruktúru DNA a v roku 1958 prvý menovaný po prvý krát vyslovil slovné spojenie *Centrálna dogma molekulárnej biológie*, ktoré hovorí, že cesta k proteínom je: DNA → RNA → proteín. Prišlo sa na to, že trojice nukleotidov DNA (kodóny) sa mapujú na aminokyseliny, prišlo sa na genetický kód (viď obrázok 2.1). Inými slovami, sekvencia nukleotidov určitého génu, ktorý definuje konkrétny proteín, sa v procese translácie „preloží“ na sekvenciu aminokyselín (podrobný popis v [3]). Dá sa povedať, že genetický kód definuje sémantiku DNA reťazca.

GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCG CCU	AGC AGU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
	UUA UUG																			
	UUA UUG																			
	UUA UUG																			
	UUA UUG																			
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Obrázok 2.1: Genetický kód (prevzaté z [3]). Pre väčšinu aminokyselín existuje viac kodónov, ktoré ich kódujú. Tri kombinácie nukleotidov sú „vyčlenené“ pre ukončovanie génov (tzv. stop-kodóny).

Spojením viacerých aminokyselín vzniká peptidový reťazec, zvyšky aminokyselín (rezi-
duá) odstupujú od osi reťazca ako tzv. postranné reťazce. O vlastnostiach proteínov roz-
hoduje charakter reziduí aminokyselín a z toho vyplývajúce sily, ktoré medzi nimi pôsobia.

Hydrofóbne (nepolárne) aminokyseliny sú priťahované k sebe, tj. do vnútra molekuly (ak sú vo vodnom prostredí), hydrofilné (polárne) sa naopak orientujú na povrch molekuly. Po denaturácii, rozpade natívnej priestorovej štruktúry, peptidového reťazca a následnom odstránení denaturačného rozpúšťadla, sa sekvencia aminokyselín zbalí späť do pôvodného tvaru. Z toho vyplýva, že úplná informácia potrebná k určeniu trojrozmerného tvaru proteínu je obsiahnutá v charaktere jeho aminokyselín a ich poradí v peptidovom reťazci. Štruktúra proteínov je pomerne zložitá, preto má zmysel definovať jej úrovne. Rozlišujeme 4 úrovne štruktúry proteínov – primárnu, sekundárnu, terciárnu a kvartérnu [3] (viď obrázok 2.2).

Primárna štruktúra. Na najnižšej úrovni, na úrovni molekúl, je sekvencia aminokyselín odvodzovaná na základe kódujúcej sekvencie nukleotidov DNA. Dlhú dobu sa sekvenovanie proteínov vykonávalo priamou analýzou ich aminokyselín. Prvým proteínom, ktorého sekvencia bola určená, je inzulín (v roku 1955). Vývoj rýchlych metód sekvenovania DNA v dnešnej dobe umožňuje oveľa jednoduchšiu, nepriamu sekvenáciu proteínov určením poradia nukleotidov v DNA [51].

Sekundárna štruktúra. Pri porovnávaní trojrozmerných štruktúr rôznych proteínov vyšlo najavo, že napriek jedinečnosti celkovej konformácie každého proteínu je v nich možné objaviť 2 základné modely skladania. Oba druhy boli objavené asi pred 60 rokmi pri štúdiu vlasov a hodvábia. Prvým z nich je α -helix (H), nájdený v proteíne α -keratín, ktorý sa hojne vyskytuje v koži, vlasoch, nechtoch atď. Druhým typom je β -sheet (E), nájdený v proteíne fibroín, ktorý je hlavnou zložkou hodvábia. Aminokyseliny mimo nich sa označujú ako Coil (C). Oba štruktúrne elementy sú stabilizované vodíkovými mostikmi. Jadra mnohých proteínov obsahujú rozsiahle oblasti β -sheetov. Tvorí sa zo susedných polypeptidových reťazcov, ktoré majú buď rovnakú alebo opačnú orientáciu, resp. sú paralelné alebo antiparalelné. α -helix vzniká, keď sa jednoduchý polypeptidový reťazec ovíja okolo samého seba a tvorí tuhý valec. Vodíkový mostík vzniká medzi každou štvrtou peptidovou väzbou a spája skupinu $C=O$ jednej peptidovej väzby so skupinou $N-H$ inej peptidovej väzby. To dáva vznik pravidelnej skrútkovici s 3,6 aminokyselinovými zvyškami (reziduami) na jednu otáčku. Krátke úseky α -helixov sú obzvlášť hojné v proteínoch umiestnených v bunčných membránach, ako sú transportné proteíny a receptory.

Terciárna štruktúra. Konečná, priestorová konformácia polypeptidového reťazca. Zisťovanie terciárnej štruktúry je metodicky veľmi zložitá, používa sa difrakcia röntgenových lúčov na kryštáloch proteínov, nukleárna magnetická rezonancia (NMR) alebo elektrónová mikroskopia [3]. Evolučne príbuzné proteíny majú veľké podobnosti v terciárnej štruktúre.

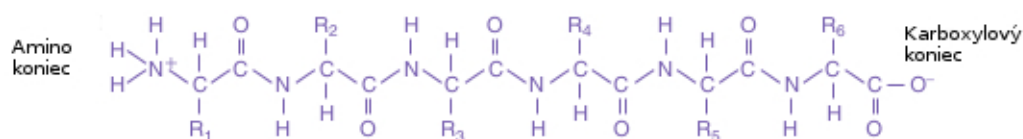
Kvartérna štruktúra. Niektoré proteíny sú zložené z väčšieho počtu menších molekúl (podjednotiek, protomérov), ktoré sú navzájom viazané nekovalentnými väzbami. Vzájomné priestorové usporiadanie týchto podjednotiek udáva kvartérnu štruktúru proteínu.

Štúdium konformácie, funkcie a evolúcie proteínov tiež prezradilo dôležitosť ďalšej organizačnej jednotky, ktorá sa líši od jednotiek doposiaľ popísaných. Touto jednotkou je *proteínová doména*, ktorá je tvorená ľubovoľnou časťou polypeptidového reťazca, ktorá sa môže nezávisle zvinúť do kompaktnej stálej štruktúry. Doména obvykle obsahuje 50–350 aminokyselín a je modulárnou jednotkou, z ktorej sú vytvorené všetky väčšie proteíny. Niekedy sa táto štruktúra nazýva *suprasekundárna*.

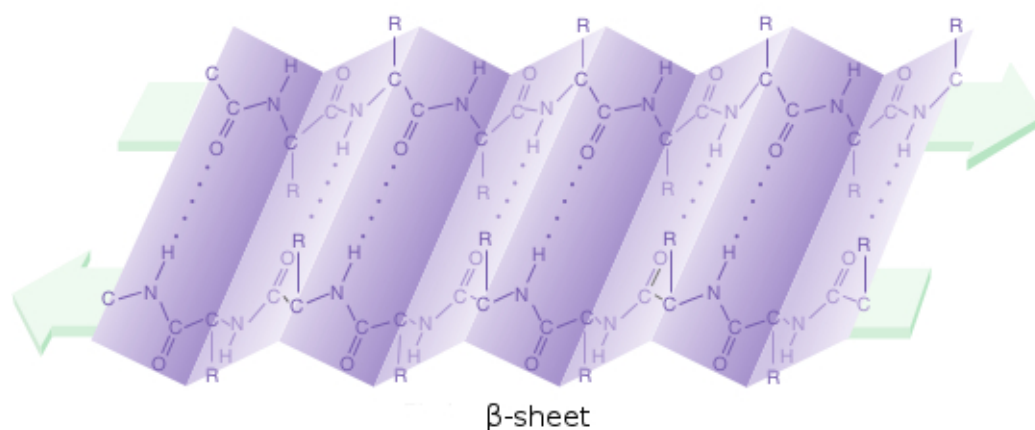
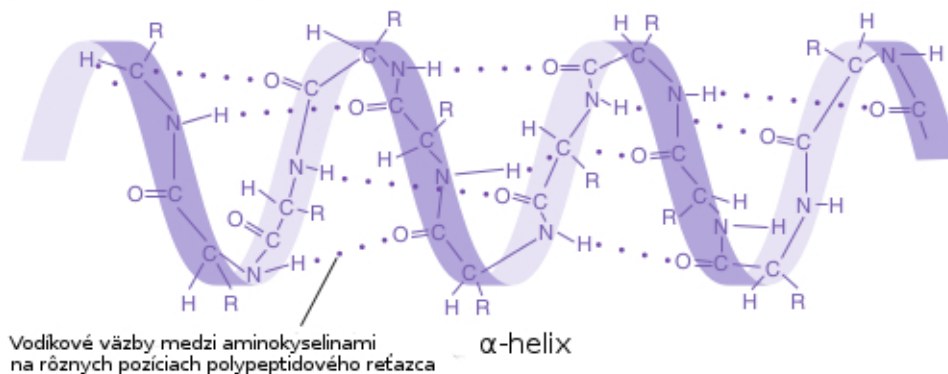
Molekuly proteínov sa zúčastňujú na všetkých základných životných procesoch. Mnohé bielkoviny sú multifunkčné, napríklad membránové imunoglobulíny imunocytov sú stavebnou súčasťou membrány a súčasne majú funkciu signálnu – rozpoznávajú „svoje“ antigény [51]. Podľa funkcie môžeme proteíny rozdeliť nasledovne:

- **Stavebné bielkoviny** – sú súčasťou bunkových štruktúr. Informácia pre špecifické

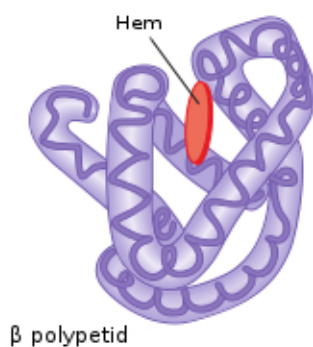
(a) Primárna štruktúra



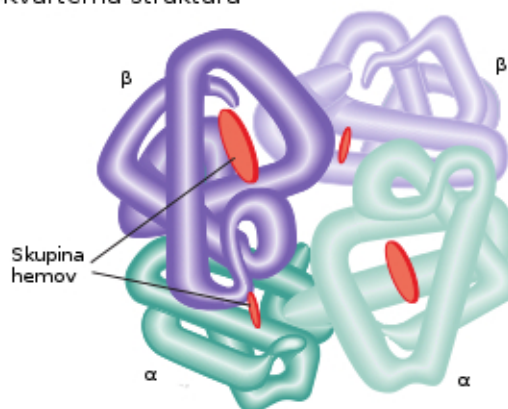
(b) Sekundárna štruktúra




(c) Terciárna štruktúra



(d) Kvartérna štruktúra




Obrázok 2.2: Úrovně štruktúry proteínu. (a) Primárna štruktúra. (b) Sekundárna štruktúra. Polypeptid môže formovať α -helix alebo β -sheet, ktorý má 2 polypeptidové segmenty usporiadané an[?]aralelne (znázornené šípkami). (c) Terciárna štruktúra. Hem je neproteínová kruhová štruktúra s atómom železe v strede. (d) Kvartérna štruktúra ilustrovaná proteínom hemoglobín, ktorý je zložený zo štyroch polypeptidových podjednotiek. Obrázok bol prevzatý z [32] a upravený.


usporiadanie podjednotiek je obsiahnutá v štruktúre molekuly, v štruktúre väzbového miesta. Nie je potrebné dodávať ani energiu, pretože nadmolekulárny komplex má nižšiu voľnú energiu ako zmes nepospájaných podjednotiek 

- **Enzýmové bielkoviny** – enzýmové reakcie uskutočňujú takmer všetky chemické reakcie v bunke, a tým celý jej metabolizmus. Enzýmová katalýza je jednou z najdôležitejších funkcií proteínov. Enzýmy umožňujú priebeh aj tých chemických reakcií, ktoré by za podmienok, v ktorých môžu živé systémy existovať, vôbec prebiehať nemohli.
- **Informačné bielkoviny** – regulujú bunkové procesy a medzibunkové vzťahy. Molekuly proteínov hrajú v týchto informačných procesoch 2 role – vystupujú ako signály, ktoré prenášajú informáciu, a receptory, ktoré môžu signály prijímať a transformovať na iné signály.

2.2 Význačné projekty súvisiace s analýzou proteínov

Potenciálu štúdia DNA, génov a ich produktov, proteínov, sú si vedomé aj vlády a každoročne investujú do výskumu množstvo finančných prostriedkov. Hlavným dotovateľom najväčších projektov je USA. V roku 1990 bol zahájený medzinárodný výskumný projekt s názvom *Projekt ľudského genómu* (HGP¹). Cieľom projektu bola sekvenácia ľudského genómu a analýza zhruba 20 000–25 000 génov z fyzikálneho aj funkčného hľadiska. V prvých fázach bol riaditeľom vyššie spomínaný James D. Watson. V roku 2003 bola publikovaná konečná verzia výsledkov a v tom istom roku bol projekt úspešne ukončený.

V roku 2011 bol ukončený projekt s názvom *Projekt 1000 genómov* (1000 Genomes Project), ktorý za pár rokov osekvenoval viac než tisíc ľudských genómov rôznych národností, zdravých aj postihnutých  účelom možnosti skúmať rôzne variácie v genóme.

Rýchlosť sekvenácie  genómu sa zrýchľuje vysokým tempom. HGP za viac než 10 rokov získal sekvenciu genómu jediného človeka, dnešné metódy, nazývané tiež Next Generation metódy, sú schopné zistiť sekvenciu genómu za rádovo dni. Taktiež cena išla nadol z miliárd na menej než \$10 000. Sekvencií dát je dostatok, no problémom je, že im príliš nerozumieme, resp. rozumieme len malej časti. Vznikla iniciatíva konkretizovaná do projektu ENCODE², ktorého cieľom je nájsť a analyzovať všetky funkčné časti ľudského genómu. Ide o rýdzo americký projekt, pracuje na ňom niekoľko pracovísk, bolo doň investovaných približne \$300 miliónov. V septembri 2012 bolo nárazovo publikovaných niekoľko desiatok prác v renomovaných vedeckých časopisoch. Jedným z výsledkov je, že nie je pravda, že väčšia časť DNA je nepotrebná, ale naopak, väčšina má určitú funkciu [56].

2.3 Predikcia sekundárnej štruktúry

Držiteľ dvoch Nobelových cien (1954, 1962), Linus Pauling, bol prvý, kto predpovedal motívy sekundárnej štruktúry proteínov (SSP) [59]. Koncom 50-tych rokov bola po prvý krát experimentálne zistená štruktúra proteínu (pomocou röntgenovej kryštalografie³). Rozkvet

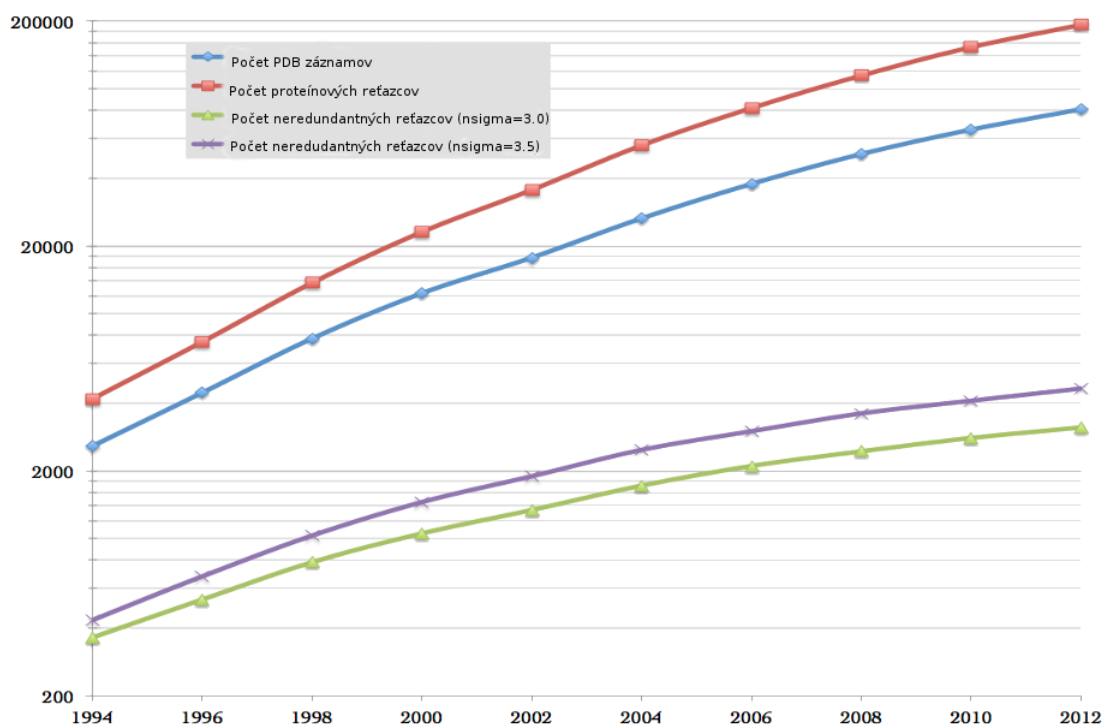
¹Z angl. Human Genome Project, domovská stránka projektu: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.

²Z angl. ENCyclopedia Of DNA Elements, domovská stránka projektu: <http://www.genome.gov/10005107>.

³Metóda zisťovania polohy jednotlivých atómov molekúl za pomoci röntgenového žiarenia, ktoré nám dovoľuje „vidieť“ rádovo v jednotkách nanometrov.

experimentálneho zisťovania štruktúry proteínov však nastal až v 90-tych rokoch 20. storočia vďaka technickému pokroku. Obrázok 2.3 ukazuje nárast počtu záznamov v súčasnosti najväčšej databáze PDB [7]. Krivky zobrazených neredundantných záznamov nemajú tendenciu konverencie, čo indikuje, že sa ešte ani neblížime úplnému pokrytiu proteínových štruktúr v rámci PDB [31]. V súčasnosti existuje približne 80 000 záznamov experimentálne zistených štruktúr proteínov, dostupná je teda databáza, na ktorú môžeme aplikovať rôzne techniky predikcie (štatistické, strojové učenie atď.).

Experimentálne metódy klasifikujú (štandardizovane podľa DSSP⁴) jednotlivé aminokyseliny do jednej z 8 tried, pri predikcii sa však v odbornej literatúre väčšinou používa redukcia na 3 základné: H, I, C \rightarrow H (Helix), E, B \rightarrow E (Beta Sheet), T, S \rightarrow C (Coil). Vyčerpávajúci popis jednotlivých tried priniesli Wolfgang Kabsch a Christian Sander v [39]. SSP problém teda znie: majme proteínovú sekvenciu s aminokyselinami $\{S_1, S_2, \dots, S_n\}$; urči pre každú S_i motív sekundárnej štruktúry – α -helix (H), β -sheet (E) alebo Coil (C). Na tento problém bolo aplikovaných veľa rôznych postupov, nasleduje klasifikácia a popis tých najúspešnejších a najprelomovejších.



Obrázok 2.3: S rokmi narastajúci počet PDB záznamov, proteínových reťazcov a neredundantných reťazcov. Úroveň redundancie sekvencií bola určovaná na základe tzv. HSSP funkcie [1]. Obrázok bol prevzatý z [34] a upravený.

Podľa chronológie možno metódy predikcie SSP rozdeliť do 3 generácií [60] (viď tabuľka 2.1). Úspešnosť metód 1. generácie nie je vysoká, čo je dané najmä neuvažovaním globálneho kontextu aminokyselinových reziduí ani evolučnej informácie extrahovanej z príslušnej rodiny homologických proteínových sekvencií. Prvé metódy sa zameriavali najmä na identifiká-

⁴Z angl. Define Secondary Structure of Proteins.

ciu α -helixov a boli založené hlavne na modeloch popisujúcich prechody medzi α -helixami a coilami [3]. Neskôr sa motív sekundárnej štruktúry pre určitú aminokyselinu určoval na základe štatistiky, ktorá uprednostňuje ten motív, ktorý je pre danú aminokyselinu najbežnejší, najpravdepodobnejší. Vtedajší nedostatok dát neumožňoval využiť plný potenciál štatistického prístupu. Medzi najvýznamnejšie techniky spadajúce do tejto generácie patrí metóda Chou-Fasman [12] a GOR (Garnier–Osguthorpe–Robson). Chou-Fasman metóda predpovedá motív sekundárnej štruktúry aktuálnej aminokyseliny na základe parametrov, ktoré vyjadrujú schopnosť predĺžiť alebo prerušiť v danom mieste motív sekundárnej štruktúry. GOR prediktor, považovaný za jedného z prvých realizovaných ako počítačový program, využíva poznatky Bayesovskej štatistiky a teórie informácie, ktoré sú aplikované na okno o veľkosti 17 aminokyselín (8 vľavo, 8 vpravo). Pre každú z 20 aminokyselín sa vypočítu frekvencie výskytu na danej pozícii v okne, na základe ktorých sa predikuje motív aminokyseliny v strede. Tento predikčný model však predpokladá, že neexistuje žiadna korelácia medzi konkrétnymi motívami sekundárnej štruktúry aminokyselín v okne 17 aminokyselín a predikovaným motívom v strede okna [27]. GOR II pracuje s rozšírenou databázou, inak je totožná so základnou metódou GOR. Tieto metódy vo vtedajšej dobe vykazovali vyššiu úspešnosť než bola reálna kvôli zahrnutiu tréningových sekvencií do testovacích [40].

Metódy vyvinuté v 80-tych rokoch 20. storočia možno považovať za 2. generáciu metód SSP. Vyšší výpočetný výkon dovoľoval zložitejšie algoritmy predikujúce motív príslušnej aminokyseliny na základe okolitých aminokyselín v definovanom okne o veľkosti 3–51 aminokyselín. Modelovala sa, na rozdiel od metódy GOR, závislosť motívu predikovanej aminokyseliny na motívoch susedných aminokyselín. Túto koreláciu si uvedomili aj tvorcovia metódy GOR, keď publikovali druhé rozšírenie – GOR III, ktoré sa považuje za najvýznamnejšieho predstaviteľa 2. generácie. Revolučiou v SSP bola dostupnosť rozsiahlych rodín homologických sekvencií. Kombinácia rozsiahlej databázy sekvencií a sofistikovaných počítačových techník viedla k prekonaniu úspešnosti 70 %.

Na prelom 2. a 3. generácie metód predikcie SSP možno zaradiť algoritmy rozšírené o ďalšie informácie o aminokyselinách, napr. tvar, veľkosť alebo fyzikálno-chemické vlastnosti. Patrí sem napríklad metóda najbližších susedov, kde sekundárna štruktúra sa určí na základe štruktúry najpodobnejších sekvení [28], GOR V [44], ZPRED [30], či PREDATOR [22], ktorý používa metódu najbližších susedov skombinovanú s interakciou so vzdialenejšími aminokyselinami.

Generácia	Obdobie	Úspešnosť [%]	Založené na
1.	1960 – 1980	50–55	predispozíciách jednotlivých aminokyselín
2.	1980 – 1990	55–62	predispozíciách segmentov aminokyselín
3.	1990 – súčasnosť	70–80	evolučnej informácii (zarovnaní sekvencií)

Tabuľka 2.1: Generácie metód predikcie SSP.

Začiatkom 90-tych rokov minulého storočia začali vznikať metódy 3. generácie, ktorých zásadnou vlastnosťou je využívanie evolučnej informácie – profilov proteínových rodín. Bolo totiž zistené, že všetky prírodne vytvorené proteíny o dĺžke viac než 100 reziduí a s viac než 35 % párovou zhodou má podobnú štruktúru, čo implikuje neskutočnú stabilitu v rámci divergencie sekvencií. Navyše, neutrálne mutácie sú veľmi nepravdepodobné, proteínov teda reálne existuje len malý zlomok, čoho dôsledkom je, úseky o dĺžke povedzme 17 reziduí implicitne obsahujú dôležité informácie o globálnych interakciách, pretože profily

viacnásobného zarovnania reflektujú evolučné obmedzenia [59]. Tieto metódy kombinujú silu väčších databáz a čoraz sofistikovanejších algoritmov založených na strojovom učení. Často sa používajú umelé neurónové siete, skryté Markovove modely, či klasifikátor SVM (z angl. Support Vector Machine). Klasifikátor SVM ukázal sa ako vhodný pre predikciu lokalizácie coilov, ktoré sú ťažko identifikovateľné štatistickými metódami [57]. Najlepšiu úspešnosť vykazujú metódy zamerané na špecifickú triedu proteínov. Medzi najznámejších predstaviteľov 3. generácie možno zaradiť PSIPRED [38], PHD [63], PHDpsi [62], PROF [61], JPred3 [15], či SSpro [58]. Táto práca sa snaží vylepšiť úspešnosť metódy *PSIPRED*, ktorá používa dvojstupňovú neurónovú sieť. Vstupom sú informácie o zarovnaní sekvencie pomocou nástroja PSI-BLAST [4]. Napriek jednoduchosti modelu je vykazovaná úspešnosť porovnateľná s ostatnými metódami 3. generácie [38].


Súčasný trend vo vývoji prediktorov SSP je vytvárať pomerne zložité modely zložené z viacerých prediktorov, ide o tzv. *konsenzuálne metódy*. Príkladom je hierarchický systém Bingru Yanga a spol., ktorý má 4 vrstvy a vykazuje úspešnosť presahujúcu 80 % [75]. Podľa štúdie z roku 2009 [5], ktorá jednotnou metodológiou analyzovala úspešnosť 59 rôznych spôsobov predikcie SSP, existujúce algoritmické techniky nemôžu byť naďalej vylepšované iba pridávaním nehomologických sekvencií do tréningovej dátovej sady, tzn. nové nástroje SSP by sa mali zamerať na navrhovanie nových techník. Dôležité je podotknúť, že nie je možné dosiahnuť úspešnosť blížiacu sa k 100 %. Teoretický horný limit úspešnosti je okolo 90 %, sčasti kvôli nejstej DSSP identifikácii blízko koncov sekundárnych štruktúr, kde sa menia lokálne konformácie [19]. Uvedená limitácia je taktiež spôsobená neschopnosťou predikcie sekundárnej štruktúry uvažovať terciárnu štruktúru. Sekvencia predikovaná ako α -helix stále môže nadobúdať konformáciu β -sheet, ak je lokalizovaná v rámci β -sheet regiónu proteínov a jeho postranné reťazce sú združené so susednými reťazcami. Lokálnu sekundárnu štruktúru taktiež môže meniť vlastná funkcia proteínu alebo aj prostredie, v ktorom sa nachádza.

Štruktúra proteínov závisí na nespočetnom množstve parametrov, ktorými sa je potrebné zaoberať, ak chceme dosahovať čoraz lepších výsledkov. Medzi tieto parametre, schopné signifikantným spôsobom zlepšiť výsledky predikcie, patrí napríklad počet kontaktov jednotlivých aminokyselín [46] alebo veľkosť chemických posunov [49].

2.4 Hodnotenie úspešnosti predikcie

Dôležitým prvkom pri vývoji metód predikcie SSP sú postupy merajúce úspešnosť týchto metód. Medzi najpoužívanéjšie úspešnostné miery patria Q_3 a SOV. Q_3 udáva pomer správne klasifikovaných reziduí proteínovej sekvencie do jednej z 3 tried (H, E, C) k všetkým reziduíam [67]. Táto metodológia je jednoduchá a má určitú výpovednú hodnotu, presne však nezachytáva „užitočnosť“ predikcie elementov sekundárnej štruktúry pre následné využitie pri predikcii terciárnej štruktúry, pretože viac než správne určenie konformačného stavu jednotlivých reziduí je dôležitejšie určenie typu a lokalizácii elementov sekundárnej štruktúry [65].

SOV (z angl. Segment Overlap) je miera, ktorá sa zameriava práve na správnu predikciu elementov sekundárnej štruktúry proteínov. Pôvodná SOV miera z roku 1994 (SOV'94) [64] nemala definovaný horný limit, čím ju nebolo možné priamo porovnávať s inými mierami (napr. s Q_3). V tejto práci používam upravenú verziu SOV (eliminujúcu nedostatky) definovanú v roku 1999 [65]. Vzhľadom k tomu, že túto mieru budem používať pri hodnotení úspešnosti SSP a jej netriviálnosti, nasledujúca časť sekcie prináša jej podrobný popis.

Nech s_1 a s_2 značia porovnávané segmenty sekundárnej štruktúry v konformačnom stave i (H, E alebo C).  segment referenčný (typicky získaný experimentálne), s_2 je segment

predikovaný. Nech (s_1, s_2) je pár prekrývajúcich sa segmentov, $S(i)$ množina všetkých prekrývajúcich sa párov segmentov v stave i a $S'(i)$ množina všetkých segmentov s_1 v stave i , pre ktoré neexistuje žiaden segment s_2 v stave i , ktorý by ich prekrýval, formálne:

$$\begin{aligned} S(i) &= \{(s_1, s_2) : s_1 \cap s_2 \neq \emptyset \wedge s_1 \text{ a } s_2 \text{ sú v konformačnom stave } i\} \\ S'(i) &= \{s_1 : \forall s_2, s_1 \cap s_2 = \emptyset \wedge s_1 \text{ a } s_2 \text{ sú v konformačnom stave } i\} \end{aligned}$$

Definícia *SOV* miery:

$$SOV = \sum_{i \in \{H, E, C\}} SOV(i) = \frac{100}{N} \sum_{i \in \{H, E, C\}} \sum_{S(i)} \left[\frac{\minov(s_1, s_2) + \delta(s_1, s_2)}{\maxov(s_1, s_2)} \times \text{len}(s_1) \right], \quad (2.1)$$

kde N je normalizačná hodnota:

$$N = \sum_{i \in \{H, E, C\}} N(i) = \sum_{i \in \{H, E, C\}} \left[\sum_{S(i)} \text{len}(s_1) + \sum_{S'(i)} \text{len}(s_1) \right], \quad (2.2)$$

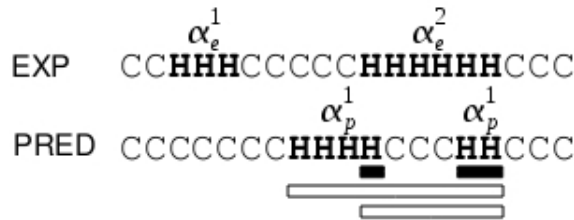
kde $\text{len}(s_1)$ vyjadruje počet reziduí v segmente s_1 , $\minov(s_1, s_2)$ dĺžku aktuálneho prekryvu segmentov s_1 a s_2 , $\maxov(s_1, s_2)$ rozsah „zjednotenia“ segmentov s_1 a s_2 a $\delta(s_1, s_2)$ je definované nasledovne:

$$\delta(s_1, s_2) = \min\{\maxov(s_1, s_2) - \minov(s_1, s_2); \minov(s_1, s_2); \lfloor \text{len}(s_1)/2 \rfloor; \lfloor \text{len}(s_2)/2 \rfloor\}, \quad (2.3)$$

kde $\min\{x_1; x_2; \dots; x_n\}$ značí minimum z n celých čísel.

Pre predstavu je uvedený príklad výpočtu *SOV* miery pre konformačný stav H pre dvojicu sekvencií zobrazenej na obrázku 2.4. Hodnota $SOV(H)$ sa na základe rovnice 2.1 vypočíta nasledovne:

$$SOV(H) = \frac{100}{6 + 6 + 3} \times \left(\frac{1 + 1}{10} + \frac{2 + 1}{6} \right) \times 6 = 28.0$$



Obrázok 2.4: Ilustrácia výpočtu $SOV(H)$. Čierne resp. biele obdĺžniky reprezentujú \minov resp. \maxov hodnoty prekrývajúcich sa segmentových párov z experimentálne zistených (EXP) a predikovaných (PRED) štruktúr.