VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

PREDIKCE SEKUNDÁRNÍ STRUKTURY PROTEINŮ POMOCÍ CELULÁRNÍCH AUTOMATŮ

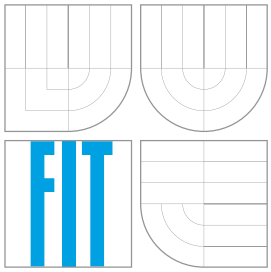
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. VLADIMÍR BRIGANT

BRNO 2013

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INFORMAČNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INFORMATION SYSTEMS

PREDIKCE SEKUNDÁRNÍ STRUKTURY PROTEINŮ POMOCÍ CELULÁRNÍCH AUTOMATŮ

PREDICTION OF SECONDARY STRUCTURE OF PROTEINS USING CELLULAR AUTOMATA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. VLADIMÍR BRIGANT

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. JAROSLAV BENDL

BRNO 2013

Abstrakt

Tato práce popisuje návrh metody predikce sekundární struktury proteinů založenou na celulárních automatech (CA). Prechodová funkce je získaná pomocí evolučního algoritmu. Predikční model bude využívat statistických i experimentálních vlastností aminokyselin. Cílem je vyvinout takovou metodu predikce, která je rychlá, vykazuje solidní úspěšnost a mohla by být použita ako doplňkový nástroj ke stávajícim metodám predikce sekunární struktury proteínů.

Abstract

This work describes a method of the secondary structure prediction of proteins based on cellular automaton (CA) model. Transition rules are acquired by evolutionary algorithm. Prediction model will use both statistical and experimental characteristics of amino acids. The goal is to create such a prediction method that is fast, performs solid and could be used as an additional tool hand in hand with currently used protein secondary structure prediction methods.

Klíčová slova

sekundární struktura proteínů, celulární automat, predikce, genetický algoritmus

Keywords

secondary protein structure, cellular automata, prediction, genetic algorithm

Citace

Vladimír Brigant: Predikce sekundární struktury proteinů pomocí celulárních automatů, diplomová práce, Brno, FIT VUT v Brně, 2013

Predikce sekundární struktury proteinů pomocí celulárních automatů

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana Ing. Jaroslava Bendla.

.....
Vladimír Brigant
9. ledna 2013

Poděkování

Chcel by som poďakovať Ing. Jaroslavovi Bendlovi za cenné rady pri tvorbe tejto práce.

Obsah

Kapitola 1

Úvod

Život na Zemi je založený na uhlíku. Chemické vlastnosti tohto prvku, ktorého tvorba by ani nezačala, keby „parametre“ vesmíru boli nastavené o trošku inak, umožňujú vytvárať dlhé uhlíkové polyméry, molekuly s uhlíkovou kostrou. Medzi tie, ktoré zabezpečujú základné funkcie života, patria nukleové kyseliny (prenos genetickej informácie), sacharidy a lipidy (zásobárne energie), a proteíny. Proteíny (viď kapitola 2) sú biopolyméry, komplexné molekuly, ich funkcia závisí na poradí aminokyselín, z ktorých sa skladajú. Význam proteínov je enormný, zabezpečujú vnútorné dýchanie, pohyb, či katalyzujú chemické reakcie.

Pokiaľ chceme pochopiť život do najmenších detailov, bez štúdia proteínov sa nezaobídeme. Podľa centrálnej dogmy molekulárnej biológie sa proteíny tvoria na základe génov v DNA. Ich funkcia je definovaná priestorovým usporiadaním jednotlivých aminokyselín (terciárna štruktúra) a naopak. Priestorové rozloženie jednotlivých aminokyselín dokážu najpresnejšie určiť experimentálne metódy. Proteínov je ale veľmi veľa (v ľudskom organizme sa odhadujú rádovo milióny) a experimentálne metódy sú časovo aj finančne náročné, čo podnietilo vznik predikčných metód štruktúry proteínov. Medzikrokok k terciárnej štruktúre proteínu je sekundárna štruktúra, ktorej správna predikcia významným spôsobom uľahčuje predikciu štruktúry priestorovej. A to je grom tejto práce – predikovať sekundárnu štruktúru proteínov. Bližší popis techník predikcie sekundárnej štruktúry proteínov sa nachádza v kapitole 3.

Použitým predikčným modelom je celulárny automat (CA) (viď kapitola 4), ktorý je tvorený mriežkou jednoduchých funkčných jednotiek (buniek). Každá bunka sa nachádza v jednom z viacerých, vopred definovaných stavov. Stav bunky sa môže počas behu CA zmeniť. Či a ako sa stav bunky zmení závisí na prechodovej funkcii CA a na stavoch buniek v okolí aktuálnej bunky. Prechodová funkcia definuje chovanie CA. Tento jednoduchý výpočetný model by mal zaistiť predovšetkým rýchlosť predikcie.

Najväčším problémom CA je vhodné určenie prechodovej funkcie. Nie je možné „odskúšať“ všetky a určiť najlepšiu, preto prichádzajú na scénu optimalizačné techniky, v našom prípade genetický algoritmus, ktorý pri hľadaní suboptimálneho riešenia používa princípy evolučného výberu a genetiky (viac v kapitole 5).

Vlastným návrhom predikčného systému (použité štatistických a experimentálnych charakteristík aminokyselín, typ CA a EA) využívajúceho uvedenú metodiku sa zaoberá kapitola 6. Záver (kapitola 7) obsahuje krátke zhrnutie práce.

Kapitola 2

Proteíny

Jacob Berzelius, významný švédsky chemik, v roku 1838 navrhol názov „proteín“ pre zložité organické zlúčeniny bohaté na dusík, ktoré nachádzal v bunkách živočíchov a rastlín. Z chemického hľadiska sú proteíny najzložitejšie a funkčne najdômyselnejšie známe molekuly [2]. Sú tvorené sekvenciou aminokyselín, ktorých väčšina bola objavená v priebehu 19. storočia a sú pospájané kovalentnou peptidovou väzbou o veľkosti približne 1,32 Å.

2.1 Tvorba proteínov

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 Crick 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 [2]). Genetický kód určuje, kedy a aké proteíny sa majú vytvoriť.

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	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
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	·

Obrázek 2.1: Genetický kód, prevzaté z [2].

2.2 Štruktúra proteínov

Pospojovaním viacerých aminokyselín vzniká peptidový reťazec, zvyšky aminokyselín odstupujú od osi reťazca ako tzv. postranné reťazce (reziduá). O vlastnostiach proteínov rozhoduje 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

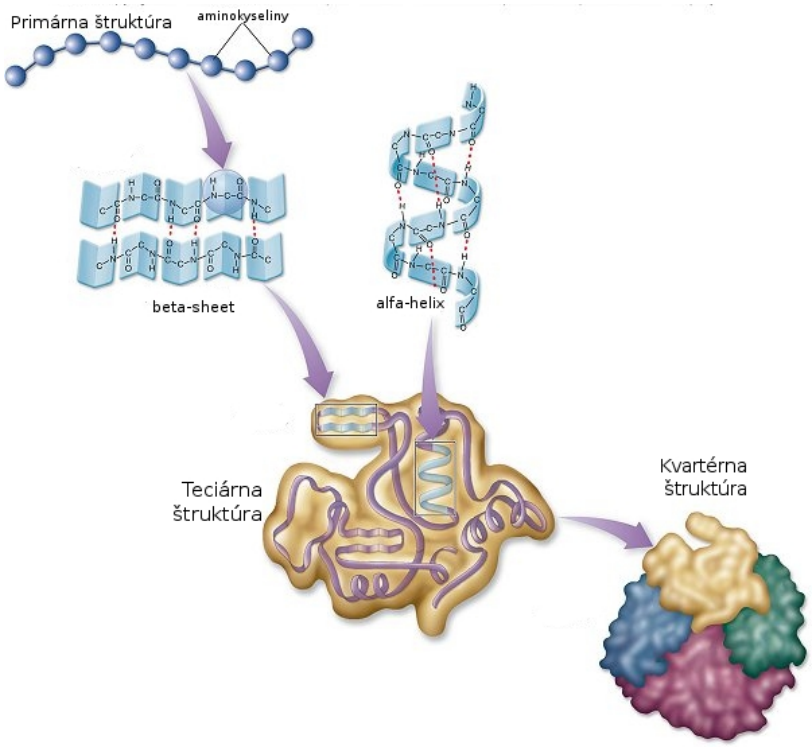
vo vodnom prostredí), hydrofilné (polárne) sa naopak orientujú na povrch molekuly. Štruktúra proteínov je pomerne zložitá, preto má zmysel definovať jej úrovne. Rozlišujeme celkovo 4 úrovne štruktúry proteínov [2] (viď obrázok 2.2):

1. **Primárna štruktúra** – sekvencia aminokyselín, ktorá sa odvodzuje podľa sekvencií nukleotidov DNA, ktoré ju kódujú (sekvenovanie DNA je metodicky jednoduchšie [22]).

2. **Sekundárna štruktúra** – časti polypeptidového reťazca, ktoré tvoria 2 základné štruktúrne pravidelnosti– α-helix (H) a β-sheet(B). Reziduá aminokyselín, ktoré nepatria do ani jedného motívu, sa označujú z historických dôvodov ako Coil (C). Ako α-helix označujeme takú konformáciu, kedy reťazec vytvára skrutkovicové usporiadanie, v β-sheet štruktúre prebiehajú úseky reťazca paralelne vedľa seba. Oba štruktúrne elementy sú stabilizované vodíkovými mostíkmi.

3. **Terciárna štruktúra** – konečná, trojrozmerná konformácia polypeptidového reťazca. Zisťovanie terciárnej štruktúry je metodicky veľmi zložité, jediným prístupom je difrakcia röntgenových lúčov na kryštáloch proteínov. Evolučne príbuzné proteíny majú veľké podobnosti v terciárnej štruktúre.

4. **Kvartérna štruktúra** – vzájomné priestorové usporiadanie podjednotiek proteínov – niektoré proteíny sú zložené z väčšieho počtu menších molekúl (podjednotiek, protomérov), ktoré sú navzájom viazané nekovalentnými väzbami.



Obrázek 2.2: Úrovne štruktúry proteínov.

2.3 Funkcie proteínov

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 [22]. Podľa funkcie môžeme proteíny rozdeliť nasledovne:

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

2. **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.

3. **Informačné bielkoviny** – regulujú bunkové procesy a medzibunkové vzťahy. Molekuly proteínov hrajú v týchto informačných procesoch 2 role – signály, ktoré prenášajú informáciu, a receptory, ktoré môžu signály prijímať a transformovať na iné signály.

2.4 Významné projekty

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 (HGP[1]). 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.

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

Rýchlosť sekvenácie genómu sa zrýchľuje vysokým tempom. HGP za viac než 10 rokov získal sekvenciu genómu 1 človeka, dnešné metódy, nazývané tiež Next Generation metódy, sú schopné zistiť sekvenciu genómu za rádovo dni, cena išla nadol z miliárd na 10-tisíce amerických dolárov (USD). Sekvencií dát je dostatok, no problémom je, že im príliš nerozumieme resp. rozumieme len malej časti. Vznikla iniciatíva konkretizovaná do projektu ENCODE[2]. Ide o rýdzo americký projekt, pracuje na ňom niekoľko pracovísk, bolo doň investovaných približne 300 miliónov USD. 5. septembra 2012 bolo 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 [24].

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

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

Kapitola 3

Predikcia sekundárnej štruktúry proteínov

Držiteľ 2 Nobelových cien (1954, 1962), Linus Pauling, bol prvý, kto predpovedal motívy sekundárnej štruktúry proteínov (SSP) [25]. Koncom 50-tych rokov bola po prvý krát experimentálne zistená štruktúra proteínu (pomocou röntgenovej kryštalografie[1]). Rozkvet experimentálneho zisťovania štruktúry proteínov však nastal až v 90-tych rokoch 20. storočia vďaka technickému pokroku, obrázok 3.1 ukazuje nárast počtu záznamov v súčasnosti najväčšej databáze PDB[2]. V súčasnosti je počet experimentálne zistených záznamov štruktúry proteínov približne 80 000, existuje teda dátová báza, na ktorú môžeme aplikovať rôzne techniky predikcie (štatistické, strojové učenie).

Experimentálne metódy klasifikujú jednotlivé aminokyseliny do 1 z 8 tried, pri predikcii sa v odbornej literatúre väčšinou používa redukcia na 3 základné: H,I,G $\rightarrow$ H (Helix), E,B $\rightarrow$ B(Beta Sheet), T,S $\rightarrow$ C(Coil). Vyčerpávajúci popis jednotlivých tried priniesli Wolfgang Kabsch a Christian Sander v [19]. SSP problém znie: pre danú proteínovú sekvenciu s aminokyselinami $\{S_1, S_2, \cdots, S_n\}$, urči pre každú S_i motív sekundárnej štruktúry – α-helix (H), β-sheet (E) alebo Coil (C). Bolo aplikovaných veľa rôznych postupov, nasleduje klasifikácia a popis tých najúspešnejších a najprelomovejších.

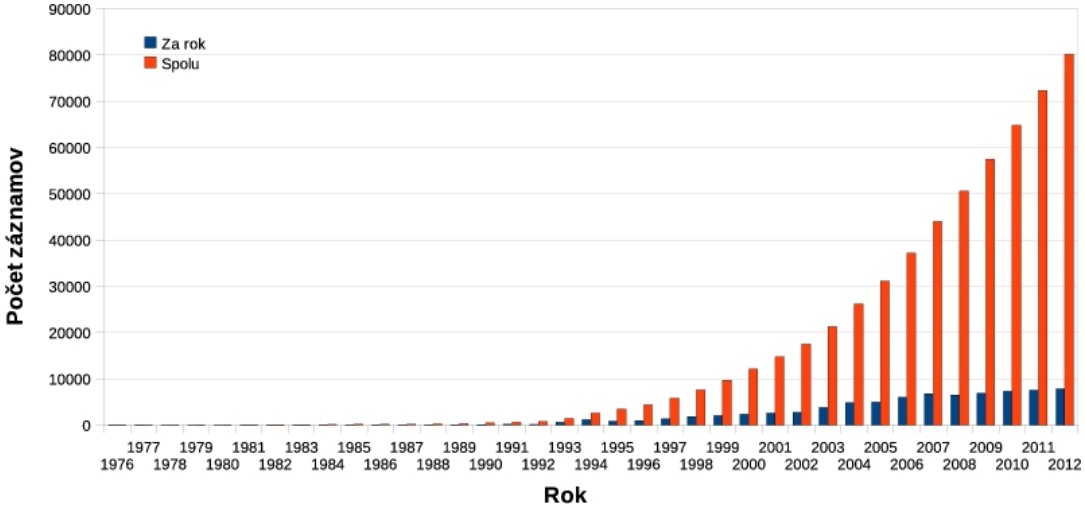
3.1 Klasifikácia predikčných metód

Navrhnutých metód predikcie bolo veľa, menej či viac úspešných. Pre lepšiu orientáciu má zmysel ich klasifikácia. Metódy predikcie SSP je možné chronologicky rozdeliť do 3 generácii [26] (viď tabuľka 3.1). Vhodné je aj rozdelenie podľa použitej predikčnej techniky (viď tabuľka 3.2). Podľa použitej predikčnej techniky delíme metódy predikcie sekundárnej štruktúry proteínov na [4]:

Spočiatku, pri nedostatočnom výkone vtedajších počítačov, sa používali metódy založené na štatistike. Ich hlavným predstaviteľom je metóda Chou-Fasman pomenovaná podľa ich autorov, Petra Y. Choua a Geralda D. Fasmana, publikovaná v roku 1974 [8]. Využíva relatívnu frekvenciu výskytu jednotlivých AK v rôznych elementoch sekundárnej štruktúry. Vyžaduje referenčnú množinu, na základe ktorej je možné vykonať štatistickú analýzu.

[1] Metóda zisťovania polohy jednotlivých atómov molekúl za pomoci röntgenového žiarenia, ktoré nám dovoluje „vidieť“ rádovo v jednotkách nanometrov.

[2] Protein Data Bank, domovská stránka: www.rcsb.org/pdb/.



Obrázek 3.1: Graf vyjadrujúci nárast počtu záznamov v databáze PDB.

Generácia	Úspešnosť [%]	Vlastnosti
1. (1960–1980)	50–55	- precenenie lokálneho kontextu na úkor globálneho - rôzne konfigurácie rovnakých aminokyselín - jednoduchosť predikcie
2. (1980–1990)	55–62	- okno - vlastnosti aminokyselín
3. (1990–2010)	70–80	- evolučná informácia

Tabulka 3.1: Generácie metód predikcie SSP.

Boli vyskúšané rôzne klasifikačné postupy, či už tie jednoduchšie, ako napríklad vzdialenostné metódy [17].

3.2 Hodnotiace prístupy

Dôležitým prvkom pri vývoji metód predikcie SSP sú postupy merajúce úspešnosť hodnotenej metódy. Medzi najpoužívanejšie patria úspešnostné miery Q_3 a SOV.

Q_3 udáva pomer správne klasifikovaných reziduí proteínovej sekvencie do 1 z 3 tried (H, E, C) k všetkým reziduám [29]. 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 [28].

Sov (z angl. segment overlap) je miera, ktorá sa zameriava na správnu predikciu elementov sekundárnej štruktúry proteínov. Pôvodná Sov miera z roku 1994 (Sov'94) [27] 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

Metódy	Generácia	Zástupci
Štatistické	I.-II.	Chou-Fasman [8], GOR
Znalostné	II.	GOR V, PREDATOR, ZPRED
Strojové učenie	II.-III.	PSIPRED [1], PHD, PROF
Konsenzuálne	III.	JPRED, NPS, CPM [35]

Tabulka 3.2: Metody predikcie

v roku 1999 [28]. Vzhľadom k tomu, že túto mieru budem používať pri hodnotení úspešnosti SSP a jej netriviálnosti, nasledujúca časť sekcie ju popisuje.

Nech s_1 a s_2 značia porovnávané segmenty sekundárnej štruktúry v konformačnom stave i (H, B alebo C). s_1 je 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 prekrývajúci sa segment s_2 v stave i, formálne:

$$S(i) = \{(s_1, s_2) : s_1 \cap s_2 \neq \varnothing \ \wedge \ s_1 \text{ a } s_2 \text{ sú v konformačnom stave } i\} \tag{3.1}$$

$$S'(i) = \{s_1 : \forall s_2, s_1 \cap s_2 = \varnothing \ \wedge \ s_1 \text{ a } s_2 \text{ sú v konformačnom stave } i\} \tag{3.2}$$

Sov miera pre konformačný stav i:

$$Sov = \sum_{i \epsilon \{H,E,C\}} Sov(i) = \frac{1}{N} \sum_{i \epsilon \{H,E,C\}} \sum_{S(i)} \left[\frac{\text{minov}(s_1, s_2) + \delta(s_1, s_2)}{\text{maxov}(s_1, s_2)} \times len(s_1) \right] \times 100 \tag{3.3}$$

kde N je normalizačná hodnota:

$$N = \sum_{i \epsilon \{H,E,C\}} N(i) = \sum_{i \epsilon \{H,E,C\}} \left[\sum_{S(i)} len(s_1) + \sum_{S(i)} len(s_1) \right] \tag{3.4}$$

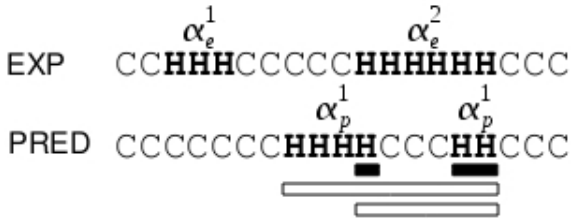
$len(s_1)$ vyjadruje počet reziduí v segmente s_1, $minov(s_1, s_2)$ dĺžku aktuálneho prekryvu segmentov s_1 a s_2, $maxov(s_1, s_2)$ rozsah „zjednotenia“ segmentov s_1 a s_2 a $\delta(s_1, s_2)$ je definované nasledovne:

$$\delta(s_1, s_2) = min\{maxov(s_1, s_2) - minov(s_1, s_2), minov(s_1, s_2), \tag{3.5}$$

$$int(len(s_1)/2), int(len(s_2)/2)\} \tag{3.6}$$

kde $min\{x_1; x_2; \ldots; x_n\}$ značí minimum z n celých čísel. Pre predstavu, na základe obrázku 3.2 a rovnice 3.3 sa hodnota Sov(H) vypočíta nasledovne:

$$Sov(H) = \frac{1}{6+6+3} \times \left(\frac{1+1}{10} + \frac{2+1}{6} \right) \times 6 \times 100 = 28.0 \tag{3.7}$$



Obrázek 3.2: 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.

Kapitola 4

Celulárne automaty

Modelovanie zložitých fyzikálnych javov pomocou počítačových simulácii sa stalo základným nástrojom pri odkrývaní tajov našej Zeme. V prírode sa stretávame s rôznymi príkladmi chovania, ktoré vykazujuje emergenciu, teda vznik vlastností systému na globálnej úrovni na základe lokálnych interakcií a bez ich explicitnej definície v rámci jednotlivých elementoch systému alebo ich prepojeniach. Ide napríklad o kolónie hmyzu, sietnicu alebo imunitný systém [30]. Jedným z cieľov umelej inteligencie je odhaliť princíp emergentného chovania ako takého. Medzi základné prístupy snažiace sa priblížiť k cieľovej páske patria agentné systémy, teória chaosu, či teória celulárnych automatov.

Koncept celulárneho automatu bol vynájdený už mnohokrát pod rôznymi názvami. V matematike je to oblasť topologickej dynamiky, v elektrotechnike iteračné polia (?), deti ich môžu poznať ako druh počítačovej hry [31]. Modelovanie pomocou celulárnych automatov je principiálne jednoduché, na základe lokálneho pôsobenia ich elementov je možné vykazovať globálne chovanie. V tejto kapitole bude model celulárneho automatu (CA) priblížený, načrtnuté historické pozadie a krátke pojednanie o jeho aplikáčných doménach.

4.1 Historické pozadie

Koncept CA uzrel svetlo sveta v 40-tych rokoch 20. storočia vďaka dvojici amerických imigrantov maďarského resp. poľského pôvodu, John von Neumannovi a Stanislawovi Ulamovi, ktorí tento koncept používali najmä pre výskum logiky života [23]. Von Neumann sa inšpiroval prácami W. McCullocha a W.Pittsa – otcov neurónových sietí [11]. Používal 2 D CA, bunky sa mohli nachádzať v 1 z 29 stavov, okolie bolo 5-bunkové (neskôr nazývané ako *von Neumannovo okolie*). Tento muž, ktorý pracoval aj na projekte Manhattan[1], dokázal existenciu konfigurácie zloženú z približne 200 000 buniek, ktorá sa dokáže samoreprodukovať tzn. takýto CA môže simulovať Turingov stroj [16]. Po von Neumannovej smrti v roku 1957 bol jeho dôkaz zjednodušovaný, za zmienku stojí Coddov celulárny automat s 8 stavmi [10], najelegantnejší dôkaz pomocou 4 stavov však priniesol v roku 1971 vo svojej dizertačnej práci Edwin Roger Banks [3]. Následovníkmi von Neumanna v štúdii celulárnych automatov boli hlavne A. Burks a jeho študent J. Holland, ktorý je však známejší z oblasti evolučných algoritmov.

V 60-tych rokoch, popri dobových, málo výkonných výpočetných zariadeniach záujem o CA ustal. O značné spopularizovanie CA sa postaral Martin Gardner, keď v roku 1970 v magazíne *Scientific American* venoval svoj stĺpček celulárnemu automatu Johna Hortona

[1] Krycí názov pre utajený americký vývoj atómovej bomby počas 2. svetovej vojny.

Conwaya s názvom „Game of Life“ [15]. Išlo o 2 D CA, ktorý je pomocou veľmi jednoduchých pravidiel schopný vykazovať, prenesene povedané, známky života.

Začiatkom 80-tych rokov sa začala skúmať otázka, či sú CA schopné modelovať okrem globálnych aspektov nášho sveta aj zákony fyziky ako také. Priekopníkmi v tomto výskume boli Tomasso Toffoli a Edward Fredkin. Hlavnou tézou ich výskumu bola definícia takých fyzikálnych výpočetných modelov, ktoré obsahujú jednu z najzákladnejších vlastností mikroskopickej fyziky – reverzibilitu. Vytvorili veľmi jednoduché modely obyčajných diferenciálnych rovníc, akými sú napríklad rovnice prúdenia tepla, vĺn, či Navier-Stokesove rovnice prúdenia tekutín [13].

Koncept CA sa postupne začal používať v rôznych oblastiach života – výpočetné úlohy, fyzikálne, chemické, biologické procesy, sociologické procesy (segregácia) atď. Obrovským prínosom pre výskum CA bola publikácia Stephena Wolframa z roku 2002 s názvom *New Kind of Science*, ktorá na 1197 stranách vyčerpávajúco analyzuje potenciál celulárnych automatov. Súčasný výpočet buniek CA dal vznik mnohým hardwarovým riešeniam ako napríklad CAM (Cellular Automata Machines) vyvinutý na MIT (Massachusetts Institute of Technology), za ktorej vývojom stoja N. Margolus a T. Toffoli.

4.2 Model CA

Všetky počítačové programy sú v princípe CA, pretože počítač pracuje s obmedzenou aritmetikou aj pamäťou. Väčšina CA však používa redukovaný stavový priestor na pár stavov (často len 2) [12].

CA je dynamický systém, v ktorom je čas aj priestor diskrétny. Skladá sa z mriežky jednoduchých funkčných jednotiek – buniek, ktoré môžu nadobúdať jeden z viacerých, vopred definovaných stavov. Stavy buniek sú synchrónne aktualizované v každom kroku výpočtu CA na základe prechodovej funkcie. Formálne [14]:

$$s^{(t+1)} = f(s^{(t)}, s_N^{(t)}), \forall i, j, \tag{4.1}$$

kde $s^{(t+1)}$ reprezentuje stav bunky danej pozíciami i a j v čase $t+1$, $s^{(t)}$ vyjadruje stav bunky v čase t, f je prechodová funkcia CA a $s_N^{(t)}$ značí stavy okolitých buniek.

Stephen Wolfrom definoval 4 triedy [33], do ktorých možno rozdeliť celulárne automaty a niektoré ďalšie výpočetné modely. Pomocou týchto tried popísal vzťah celulárnych automatov k dynamickým systémom (uvedené v zátvorkých):

- **Trieda I** – počiatočné konfigurácie evolvujú do stabilného, homogénneho stavu. Akákoľvek náhodnosť počiatočnej konfigurácie mizne (limitné body).
- **Trieda II** – počiatočné konfigurácie konvergujú do jednoducho separovateľných periodických štruktúr (limitné cykly).
- **Trieda III** – vykazuje chaotické neperiodické vzory (chaotické chovanie podobné podivným atraktorom).
- **Trieda IV** – vykazuje komplexné vzory (veľmi dlhé prechodné úseky, ktoré nemajú jasnú analógiu v spojitých dynamických systémoch).

Medzi najznámejšie príklady celulárnych automatov patria:

- **Game of Life** – je spomenutý v každej spisbe týkajúcej sa CA, pripomína vývoj spoločenstva živých organizmov. Existuje veľké množstvo implementácii[2]. Každá bunka sa nachádza v 1 z 2 stavov – mrtvá alebo živá. Okolie je 5-bunkové, von Neumannove. Pravidlá hry resp. prechodová funkcia je veľmi jednoduchá:
  - bunky s menej než 2 živými susedmi zomrú
  - živé bunky s 2 alebo 3 živými susedmi prežívajú
  - živé bunky s viac než 3 živými bunkami zomrú
  - mrtvé bunky s práve 3 živými susedmi ožijú
- **Coddov automat** –
- **Langtonove Q-slučky** – neporovnateľne jednoduchšia verzia na báze Coddovho automatu
- **CAM** –

4.3 Modifikácie modelu CA

V praktických experimentoch sa klasický model CA, popísaný v sekcii 4.2, používa len občas. Väčšinou sa vyskytuje v rôznych modifikáciach.

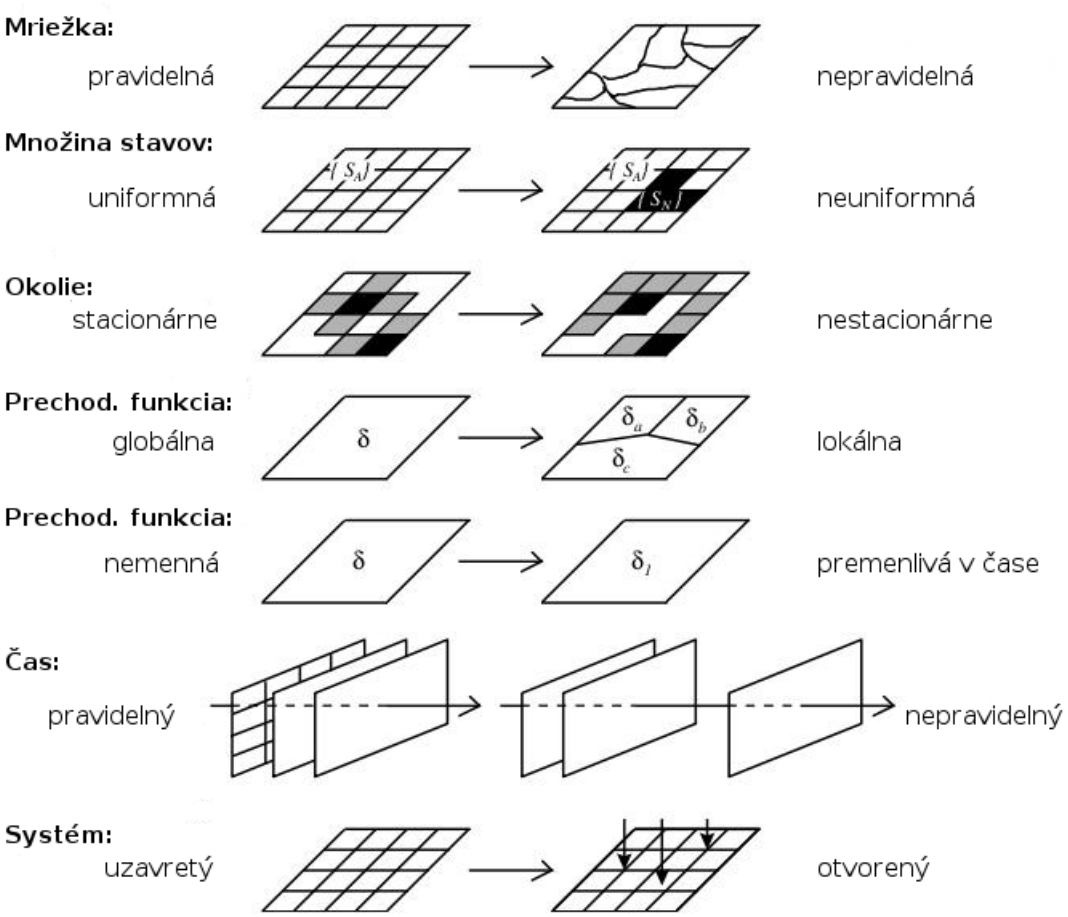
4.4 Aplikácie modelu CA

Významné postavenie má CA v modelovaní biochemických javov, príkladom je predikcia miesta pôsobenia proteínov v bunke. Cieľom je vytvoriť automatizovanú metódu spoľahlivého určovania polohy skúmaných proteínov. Informácia o polohe proteínu dokáže výrazne urýchliť proces určovania jeho biologickej funkcie. CA sa v tomto prípade používa na tvorbu „obrázkov“, na ktoré sa aplikujú metódy rozpoznávania vzorov v obraze [34].

Model „bunkového“ automatu možno použiť aj na modelovanie biologických dráh, konkrétne na modelovanie signálnych dráh mitogénom aktivovaných proteínkynáz [20]. Ide o signálnu dráhu, po ktorej sú signály vysielané z cytoplazmatickej membrány do cytoplazmy a jadra. Použitý CA modeluje 3 substráty a 4 enzýmy. Koncentrácie jednotlivých substrátov a enzýmov sú definované ako počet buniek CA.

Ermentrout a Edelstein-Keshet spísali prehľadový článok aplikácii modelu CA v modelovaní biologických procesov [12], v ktorom popísali možnosť použitia modelu CA v ...

[2] Interaktívnu implementáciu v podobe appletu možno nájsť na adrese: http://www.bitstorm.org/gameoflife/.



Obrázek 4.1: Klasický vs. modifikovaný CA, upravené a prevzaté z [6].

Kapitola 5

Evolučné algoritmy

Charles Darwin bol angličan, syn významného lekára. Vyštudoval teológiu, po štúdiu sa zaoberal geologickými formáciami v horách Walesu. Koncom roka 1831 však odišiel na 5-ročnú výskumnú cestu okolo sveta. Loď HMS Beagle ho zaviedla aj na Galapágy, ekvádorské súostrovie 19 sopečných ostrovov vo východnej časti Tichého oceánu, kde zhromaždil podľa jeho slov najcennejšiu časť prírodovedeckého materiálu, ktorý použil vo svojom najväčšom diele publikovanom v roku 1859 s názvom *O vzniku druhov prírodným výberom alebo uchovávanie prospešných plemien v boji o život*. Vrhá ucelený pohľad na vývoj druhov oslobodený od spirituality a náboženských predstáv svojej a predchádzajúcich dôb a hlavou plnou prírodovedných informácii získaných z okružnej plavby okolo sveta. Darwin vysvetľuje vznik rôznych druhov organizmov na základe prirodzeného výberu, teda schopnosti prežiť len tých najschopnejších. Významným argumentom pre jeho teóriu boli aj stratigrafické výsledky geológa Charlesa Lyella, ktoré podporovali rodiacu sa evolučnú teóriu v „časovej zložitosti“. Aplikované princípy klasickej evolučnej teórie s mnohými „vylepšeniami“ hrajú významnú úlohu vo fonde vedomostí ľudstva.

Evolučné algoritmy, ktoré sú postavené na myšlienkach evolučnej teórie, začali vznikať už v 50-tych rokoch 20. storočia. Výraznejší záujem však nastal až v 80-tych rokoch (?[18])...

5.1 Biologické pojmy v EA kontexte

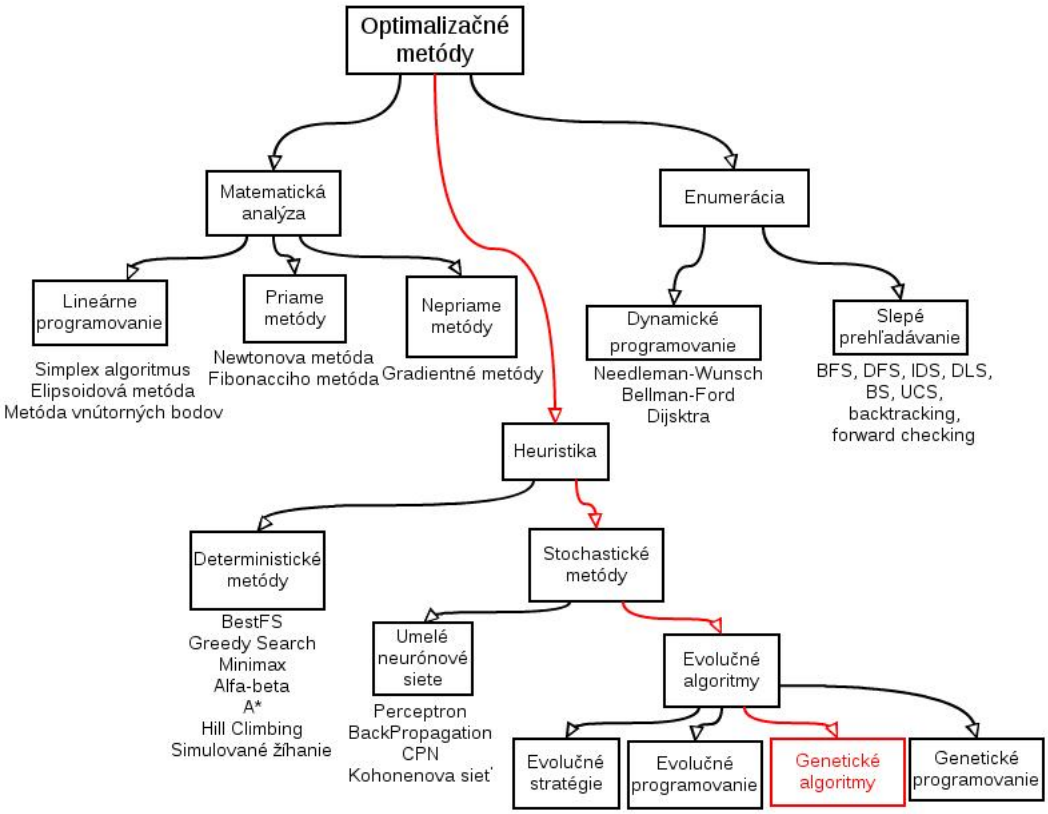
Vo zvyšku tejto práce sa budú vyskytovať pojmy pochádzajúce z biológie, no sémantika nie všetkých je zhodná so sémantikou v kontexte evolučných algoritmov. Pre ujasnenie pojmov ponúkam ich krátky prehľad.

5.2 Zaradenie evolučných algoritmov

EA 5.1

5.3 Genetické operátory

Evolučné procesy z biológie boli aplikované a svojím spôsobom interpretované v teórii evolučných algoritmov. Ide o pekný príklad medzioborového transféru informácií. Genetickými operátormi sú selekcia (prirodzený výber) a rekombinačné operátory – kríženie a mutácia.



Obrázek 5.1: Klasifikácia genetických algoritmov v kontexte optimalizačných techník.

5.3.1 Selekcia

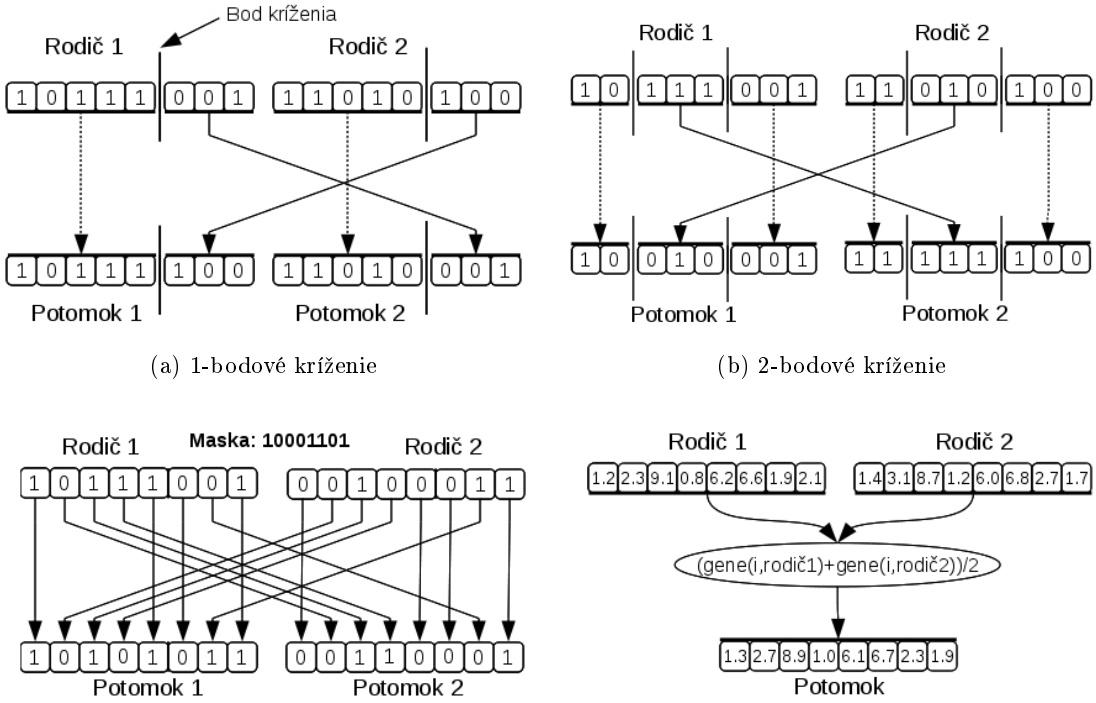
...

5.3.2 Kríženie

...

5.3.3 Mutácia

...



(a) 1-bodové kríženie

(b) 2-bodové kríženie

(c) Kríženie maskou

(d) Aritmetické kríženie (pomocou priemeru)

Obrázek 5.2: Rôzne typy evolučného operátoru kríženia, prevzaté z [5].

Kapitola 6

Návrh predikčného systému

Navrhnutý systém z veľkej časti vychádza z 2 prác venujúcich sa SSP pomocou CA [7] [32]. Modifikované sú tie časti systému, ktoré majú potenciál zlepšiť úspešnosť predikcie SSP. Ide hlavne o spôsob klasifikácie jednotlivých reziduí popísaný nižšie. Aby sa predikčný model mal o čo oprieť, definujeme štatistické (sekcia 6.1) a experimentálne (sekcia 6.2) vlastnosti reziduí.

6.1 Štatistický popis reziduí

Štatistika je mocný nástroj, veda, pomocou ktorej na základe empirických dát získavame nové znalosti. V návrhu systému budú využité 2 štatistické vlastnosti aminokyselín, Chou–Fasmanove parametre, ktoré aminokyseliny charakterizujú z pohľadu miery výskytu v určitom motíve sekundárnej štruktúry, a konformačná analýza, ktorá klasifikuje jednotlivé aminokyseliny pre konkrétny motív sekundárnej štruktúry do určitej konformačnej triedy na základe toho, v akej časti motívu sekundárnej štruktúry sa príslušná aminokyselina nachádza.

6.1.1 Chou-Fasmanove parametre

Jednou z metód predikcie SSP 1. generácie je Chou-Fasmanova metóda, bližšie popísaná v sekcii 3.1. V ich práci z roku 1974 [9] je definovaný tzv. parameter konformačnej preferencie aminokyseliny j ku konformačnému stavu i, Chou-Fasmanov parameter P_j^i:

$$P_j^i = \frac{f_j^i}{\langle f_j^i \rangle} \tag{6.1}$$

kde f_j^i je relatívna frekvencia aminokyseliny j v konformačnom stave i daná vzťahom 6.2 a $\langle f_j^i \rangle$ priemerná relatívna frekvencia konformačného stavu i v rámci všetkých aminokyselín vyjadrená vzťahom 6.3.

$$f_j^i = \frac{n_j^i}{n^i} \tag{6.2}$$

kde n_j^i je počet reziduí j v konformačnom stave i a n^i je celkový počet reziduí v konformačnom stave i.

$$\langle f_j^i \rangle = \frac{\sum_{\forall k \, \epsilon \, AK} f_k^i}{n_j} \tag{6.3}$$

6.1.2 Konformačné triedy

Na základe práce Guang-Zheng Zhanga a spol. [36] definujeme konformačnú triedu pre všetky aminokyseliny a všetky konformačné stavy (H, B, C). Nech $P = p_1, p_2, \ldots, p_n$ je primárna štruktúra proteínu (sekvencia aminokyselín) a $S = s_1, s_2, \ldots, p_n$ odpovedajúca sekundárna štruktúra proteínu dĺžky n. Ak s_i a s_{i+1} sú rôzne konformačné stavy, napríklad $s_i = H$ a $s_{i+1} = B$, hovoríme o tzv. štruktúrnom prechode (ST[1]), v tomto prípade ST_{HB}. Týmto spôsobom môžeme definovať ostatných 5 štruktúrnych prechodov: ST_{HC}, ST_{BH}, ST_{BC}, ST_{CH} a ST_{CB}.

Na základe uvedených štruktúrnych prechodov definujeme konformačnú preferenciu aminokyselín pre určitý motív sekundárnej štruktúry. Štruktúrny prechod ST_{HB} môžeme chápať ako ukončenie H a súčasne ako začiatok B. V kontexte všetkých 6 ST, počet všetkých ukončení a začiatkov H určíme nasledovne:

$$N_{\alpha\text{-ukončenie}} = N_{ST_{HB}} + N_{ST_{HC}} \tag{6.4}$$

$$N_{\alpha\text{-začiatok}} = N_{ST_{BH}} + N_{ST_{CH}} \tag{6.5}$$

kde $N(\cdot)$ reprezentuje počet rôznych štruktúrnych prechodov. Počet ukončení a začiatkov B a C sa určí analogicky. Definujeme konformačnú preferenciu CP ukončenia resp. začiatku určitého konformačného stavu aminokyseliny i, konkrétne $CP_{j,\alpha\text{-ukončenie}}$, $CP_{j,\alpha\text{-začiatok}}$, $CP_{j,\beta\text{-ukončenie}}$, $CP_{j,\beta\text{-začiatok}}$, $CP_{j,\text{Coil}-\text{ukončenie}}$ a $CP_{j,\text{Coil-začiatok}}$. Výpočet $CP_{j,\alpha\text{-ukončenie}}$ (ostatné konformačné preferencie sa získajú analogicky):

$$CP_{j,\alpha\text{-ukončenie}} = \frac{P_{j,\alpha\text{-ukončenie}}}{P_j} \tag{6.6}$$

kde $P_{j,\alpha\text{-ukončenie}}$ a P_j sa získa nasledovne:

$$P_{j,\alpha\text{-ukončenie}} = \frac{N_{j,\alpha\text{-ukončenie}}}{\sum_{i=1}^{20} N_{i,\alpha\text{-ukončenie}}} \tag{6.7}$$

kde $N_{j,\alpha\text{-ukončenie}}$ vyjadruje počet reziduí ukončujúcich H.

$$P_j = \frac{N_j}{N} \tag{6.8}$$

kde N_j vyjadruje celkový počet reziduí aminokyseliny j a N celkový počet reziduí. Uvažujúc reziduum j a jeho motív sekundárnej štruktúry, α-helix (H), definujeme konformačnú triedu (CC) konformačného stavu rezidua j (obdobne možno vyjadriť konformačné triedy pre B a C):

[1] Z angl. Structure Transition.

$$CC_{j,\alpha} = \begin{cases} b & \text{ak } CP_{j,\alpha\text{-ukončenie}} \geq 1 \ \wedge \ CP_{j,\alpha\text{-začiatok}} < 1 \\ f & \text{ak } CP_{j,\alpha\text{-začiatok}} \geq 1 \ \wedge \ CP_{j,\alpha\text{-ukončenie}} < 1 \\ n & \text{inak} \end{cases} \tag{6.9}$$

Použité znaky b, f, n značia v tomto poradí triedy *Breaker*, *Former* a *Neutral*. Konformačná klasifikácia rezidua j je vyjadrená 3-znakovým kódom – $CC_{j,\alpha}CC_{j,\beta}CC_{j,Coil}$.

6.2 Experimentálny popis reziduí

5.1.2013 bol publikovaný článok, ktorý predikuje sek. štruktúru na základe chemických „shiftov“ [21], existuje databáza VASCO, čo by vyžadovalo BLAST + pick do VASCO, to do SEP dávať nebudem, keď budem mať ostatné experimenty hotové, môžem vyskúšať aj so shiftami.

6.3 Použitý celulárny automat

Sekvenciu aminokyselín proteínov reprezentuje 1 D CA, ktorého bunky modelujú jednotlivé reziduá aminokyselín. Bunky môžu nadobúdať 1 z 3 stavov (H,B,C). Veľkosť okolia nie je pevná, evolučný algoritmus zisťuje optimálny rádius. Štatistické a experimentálne vlastnosti aminokyselín sú modelované parametrami buniek CA. V rámci inicializácie sú každej bunke pridelené Chou-Fasmanove koeficienty, ktoré sú pri prechode do ďalšej konfigurácie CA upravované. Aktualizovaná je aj konformačná trieda buniek na základe tabuľky definovanej pre konkrétnu dátovú sadu.

6.3.1 Prechodová funkcia

Myšlienkou prechodovej funkcie je výpočet preferencie bunky/aminokyseliny výskytu v určitom konformačnom stave na základe príslušnej konformačnej triedy CC_j^i a hodnôt parametrov vychádzajúcich z Chou-Fasmanových parametrov P_j^i, CF parametrov:

$$P_{t+1,j}^i = \frac{\sum_{k=-o}^{o} w_k P_{j-k}^i}{\sum_{k=-o}^{o} w_k} \tag{6.10}$$

$P_{t+1,j}^i$ vyjadruje CF parameter bunky j v čase $t+1$ pre konformačný stav i, ktorý je váhovaným súčtom jednotlivých CF parametrov P_{j-k}^i v okolí o. Vlastná prechodová funkcia má tvar:

$$S_{t+1,j} = \max R_{t+1,j}^i \quad i \,\epsilon\, \{H, B, C\} \tag{6.11}$$

kde $S_{t+1,j}$ je stav bunky j v čase $t+1$ a parameter $R_{t+1,j}^i$ vyjadruje mieru príslušnosti bunky resp. aminokyseliny j v kroku $t+1$ ku konformačnému stavu i (H, B, C):

$$R_{t+1,j}^i = P_{t+1,j}^i + \alpha \cdot CC_j^i \tag{6.12}$$

kde α je koeficient miery závislosti stavu bunky na konformačnej triede CC_j^i optimalizovaná evolučným algoritmom.

6.4 Použitý evolučný algoritmus

EA, konkrétne evolučná stratégia (ES), je aplikovaná na optimalizáciu prechodovej funkcie CA. Evolvovaný chromozóm má tvar:

$$C = [s, \alpha, r, w_{-r}, w_{-r+1}, \ldots, w_{r-1}, w_r] \tag{6.13}$$

kde s vyjadruje počet krokov EA, α signifikantnosť konformačnej triedy na stav bunky a teda na úspešnosť predikčného modelu, r veľkosť okolia a w_i pre $i \epsilon \{-r, \ldots, r\}$ váhy jednotlivých buniek okolia.

Kapitola 7

Záver

Proteíny sú základnými stavebnými kameňmi života na Zemi. Okrem využitia v posilňovni sa starajú o podstatnú časť biologických funkcí a ich reguláciu. Funkcia proteínov je určená ich štruktúrou a predikciou (sekundárnej) štruktúry sa venovala táto práca. Bol navrhnutý predikčný model založený na modeli celulárneho automatu, ktorý má viacero stupňov volnosti – počet krokov automatu, veľkosť okolia, či váhy jednotlivých buniek v okolí. Kvôli netriviálnej optimalizácii týchto parametrov boli využité služby evolučných algoritmov. Navrhnutý systém si kladie za priority rýchlosť a solídnu úspešnosť, ktorou nemôže konkurovať sofistikovaným predikčným metódam, ale môže slúžiť ako doplnkový nástroj.

Literatura

[1] The PSIPRED protein structure prediction server. *Bioinformatics*, ročník 16, č. 4, Duben 2000: s. 404–405, ISSN 1460-2059.

[2] Alberts, B.; Bray, D.; Johnson, A.; aj.: *Základy buněčné biologie: úvod do molekulární biologie buňky*. Ústí nad Labem: Espero Publishing s.r.o, druhé vydání, 2005, ISBN 80-902906-2-0, 740 s.

[3] Banks, E. R.: Information processing and transmission in cellular automata. Technická zpráva, Cambridge, MA, USA, 1971.

[4] Bendl, J.; Burgetová, I.: Proteínové predikce. University Lecture, 2011, [online],[cit. 2012-12-05].

[5] Brigant, V.: *Evoluční návrh simulátoru založeného na celulárních automatech.* Bakalářská práce, FIT VUT v Brně, Brno, 2011.

[6] Cecchini, A.; Rinaldi, E.: The multi - cellular automaton: a tool to build more sophisticated models. A theoretical foundation and a practical implementation. 1999.

[7] Chopra, P.; Bender, A.: Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature. *In Silico Biol*, ročník 7, č. 1, 2007: s. 87–93.

[8] Chou, P. Y.; ; Fasman, G. D.: Prediction of protein conformation. *Biochemistry*, ročník 13, č. 2, Leden 1974: s. 222–245.

[9] Chou, P. Y.; Fasman, G. D.: Conformational parameters for amino acids in helical, Î¸-sheet, and random coil regions calculated from proteins. *Biochemistry*, ročník 13, č. 2, Leden 1974: s. 211–222.

[10] Codd, E. F.: *Cellular Automata*. Orlando, FL, USA: Academic Press, Inc., 1968, ISBN 0121788504.

[11] Delorme, M.; Mazoyer, J.: *Cellular Automata: a parallel model*, ročník 460. Springer, 1998.

[12] Ermentrout, B. G.; Edelstein-Keshet, L.: Cellular Automata Approaches to Biological Modeling. *Journal of Theoretical Biology*, ročník 160, č. 1, Leden 1993: s. 97–133, ISSN 00225193.

[13] Fredkin, E.; Toffoli, T.: Collision-based computing. kapitola Conservative logic, London, UK, UK: Springer-Verlag, 2002, ISBN 1-85233-540-8, s. 47–81.

[14] Fuqiang, D.: Mining Dynamic Transition Rules of Cellular Automata in Urban Population Simulation. In *Proceedings of the 2010 Second International Conference on Computer Modeling and Simulation - Volume 02*, ICCMS '10, Washington, DC, USA: IEEE Computer Society, 2010, ISBN 978-0-7695-3941-6, s. 471–474.

[15] Gardner, M.: Mathematical Games The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, ročník 223, 1970: s. 120–123.

[16] Gardner, M.: *Wheels, life, and other mathematical amusements*. Freeman, 1983.

[17] Ghosh, A.; Parai, B.: Protein secondary structure prediction using distance based classifiers. *Int. J. Approx. Reasoning*, ročník 47, č. 1, Leden 2008: s. 37–44, ISSN 0888-613X.

[18] Hynek, J.: *Genetické algoritmy a genetické programování*. Praha 7: Grada Publishing a.s., 2008, ISBN 978-80-247-2695-3.

[19] Kabsch, W.; Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, ročník 22, č. 12, Prosinec 1983: s. 2577–2637, ISSN 0006-3525.

[20] Kier, L. B.; Bonchev, D.; Buck, G. A.: Modeling Biochemical Networks: A Cellular-Automata Approach. *Chemistry & Biodiversity*, ročník 2, č. 2, 2005: s. 233–243, ISSN 1612-1880.

[21] Mechelke, M.; Habeck, M.: A probabilistic model for secondary structure prediction from protein chemical shifts. *Proteins: Structure, Function, and Bioinformatics*, 2012, ISSN 1097-0134.

[22] Nečas, O.: *Obecná biologie pro lékařské fakulty*. H & H Vyšehradská, 2000, ISBN 9788086022468.

[23] von Neumann, J.: *Theory of Self-Reproducing Automata*, ročník 160. Illinois: University of Illinois Press, 1966.

[24] Pennisi, E.: Genomics. ENCODE project writes eulogy for junk DNA. *Science*, ročník 337, č. 6099, 2012: s. 1159, 1161, ISSN 1095-9203.

[25] Rost, B.: Review: Protein Secondary Structure Prediction Continues to Rise. *J. Struct. Biol*, ročník 134, 2001: s. 204–218.

[26] Rost, B.: Protein Prediction - Part 1: Structure. University Lecture, 2011, [online],[cit. 2012-12-05].

[27] Rost, B.; Sander, C.; Schneider, R.: Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology*, ročník 235, č. 1, 1994: s. 13 – 26, ISSN 0022-2836.

[28] Rost, B.; Zemla, A.; Fidelis, K.; aj.: A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins: Structure, Function, and Genetics*, ročník 34, 1999: s. 220–223.

[29] Schulz, G. E.; Pain, R. H.; Schirmer, R. H.: Principles of protein structure. *Biochemical Education*, ročník 8, č. 4, 1980: s. 108–130, ISSN 1879-1468.

[30] Sipper, M.: *Evolution of Parallel Cellular Machines: The Cellular Programming Approach*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2001, ISBN 3540626131.

[31] Toffoli, T.; Margolus, N.: *Cellular automata Machines, A New Environment For Modeling*. Cambridge, MA: MIT Press, 1987.

[32] Šalanda, V.: *Predikce sekundární struktury proteinu pomocí celulárního automatu*. Bakalářská práce, FIT VUT v Brně, Brno, 2012.

[33] Wolfram, S.: *A New Kind of Science*. Wolfram Media, January 2002, ISBN 1-57955-008-8, 1197 s.

[34] Xiao, X.; Shao, S.; Ding, Y.; aj.: Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids*, ročník 30, 2006: s. 49–54, ISSN 0939-4451.

[35] Yang, B.; Hou, W.; Xie, Y.; aj.: The Research of Protein Secondary Structure Prediction System Based on KDTICM. *Proceedings of The World Congress on Engineering and Computer Science*, 2009: s. 47–51.

[36] Zhang, G.-Z.; Huang, D. S.; Zhu, Y. P.; aj.: Improving protein secondary structure prediction by using the residue conformational classes. *Pattern Recogn. Lett.*, ročník 26, č. 15, Listopad 2005: s. 2346–2352, ISSN 0167-8655.