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

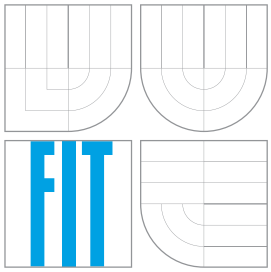
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

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 sekundární 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, reasonably accurate and could be used as an additional tool hand in hand with today's used protein secondary structure prediction methods.

Klíčová slova

sekundární struktura proteínů, celulární automat, proteinové predikce, genetický algoritmus

Keywords

secondary protein structure, cellular automata, protein 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
11. května 2013

Poděkování

Velké poděkování takisto patří za přístup k výpočetním a úložním zařízením patřícím do národní síťové infrastruktury MetaCentrum, poskytovaným v rámci programu „Projekty velké infrastruktury pro výzkum, vývoj a inovaci“ (LM2010005).

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. Toto usporiadanie dokážu najpresnejšie (nie však na 100 %) určiť experimentálne metódy. Proteínov je však veľmi veľa (v ľudskom organizme sa ich počet odhaduje na rádovo milióny [43]) a neefektívnosť časovo a finančne náročných experimentálnych metód podnietila vznik strojových predikčných metód štruktúry proteínov. Medzikrokom k terciárnej štruktúre proteínu je sekundárna štruktúra, ktorej správna predikcia významným spôsobom uľahčuje predikciu štruktúry priestorovej. A to je grom tejto práce – vylepšiť predikciu sekundárnej štruktúry proteínov. Bližší popis základných 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 meniť. Či a ako sa stav bunky zmení, závisí na prechodovej funkcii CA a na stavoch buniek v okolí aktuálnej bunky. Prechodová funkcia definuje správanie CA. Tento jednoduchý výpočetný model by mal zaistiť predovšetkým rýchlosť predikcie.

Najväčším problémom CA je vhodné určenie prechodovej funkcie. Nie je možné „odskúšať“ všetky možné a určiť najlepšiu z nich, 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, použitím štatistických a experimentálnych charakteristík aminokyselín, celulárneho automatu a evolučne inšpirovaných techník, sa zaoberá kapitola 6.

V kapitole 7 sa nachádza popis implementácie konzolovej aj webovej verzie systému. Popis a spracovanie experimentov (optimálna veľkosť okolia, počet krokov CA či predikcia v spolupráci s nástrojom PSIPRED) obsahuje kapitola 8. Záver (kapitola 9) 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 [3]. 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 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]). 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 [3].

2.2 Štruktúra a funkcia proteínov

Pospájaním 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

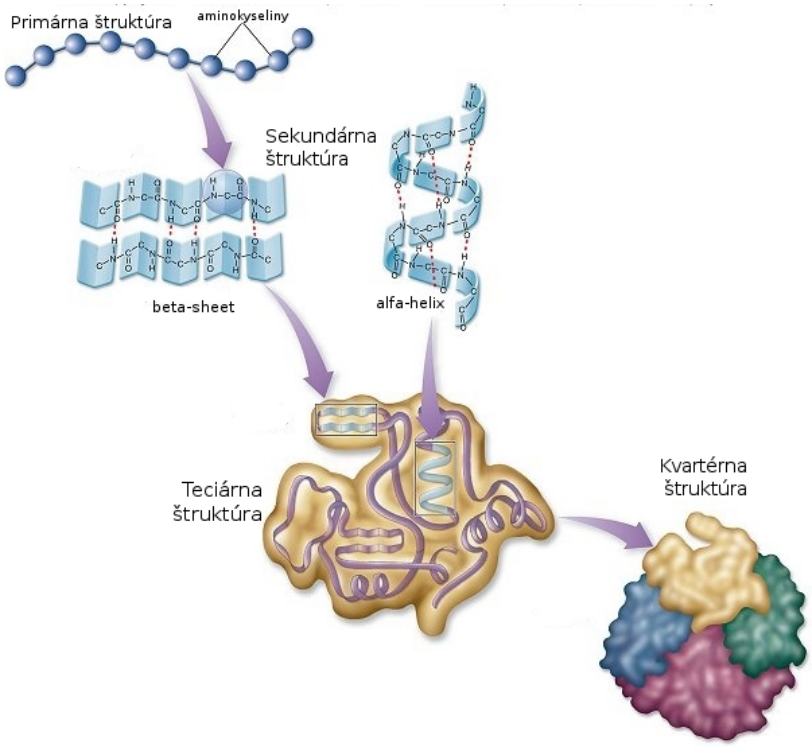
vo vodnom prostredí), hydrofilné (polárne) sa naopak orientujú na povrch molekuly. Štruktúra proteínov je pomerne zložitá, preto má zmysel definovať jej úrovne. Rozlišujeme celkovo 4 úrovne štruktúry proteínov [3] (viď obrázok 2.2):

1. **Primárna štruktúra** – sekvencia aminokyselín odvodzovaná podľa kódujúcej sekvencie nukleotidov DNA (sekvenovanie DNA je metodicky jednoduchšie [46]).

2. **Sekundárna štruktúra** – časti polypeptidového reťazca, ktoré tvoria 2 základné štruktúrne pravidelnosti – α-helix (H) a β-sheet (E). Reziduá aminokyselín, ktoré nepatria ani do 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 alebo antiparalelne vedľa seba. Oba štruktúrne elementy sú stabilizované vodíkovými mostíkmi.

3. **Terciárna štruktúra** – konečná, trojrozmerná konformácia polypeptidového reťazca. Zisťovanie terciárnej štruktúry je metodicky veľmi zložité, používa sa difrakcia röntgenových lúčov na kryštáloch proteínov, nukleárna magnetická rezonancia (NMR) alebo elektrónová mikroskopia. 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, prevzaté z [1].

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 [46]. 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.

2.3 Významné 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 (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 jediného človeka, dnešné metódy, nazývané tiež Next Generation metódy, sú schopné zistiť sekvenciu genómu za rádovo dni, 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 ENCODE[2], 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 USD. 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 [49].

[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 dvoch Nobelových cien (1954, 1962), Linus Pauling, bol prvý, kto predpovedal motívy sekundárnej štruktúry proteínov (SSP) [52]. 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 [6]. 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 DSSP[2]) jednotlivé aminokyseliny do 1 z 8 tried, pri predikcii sa však v odbornej literatúre väčšinou používa redukcia na 3 základné: H, I, G $\rightarrow$ H (Helix), E, B $\rightarrow$ E (Beta Sheet), T, S $\rightarrow$ C (Coil). Vyčerpávajúci popis jednotlivých tried priniesli Wolfgang Kabsch a Christian Sander v [35]. SSP problém teda znie: majme 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). Na tento problém bolo aplikovaných veľa rôznych postupov, nasleduje klasifikácia a popis tých najúspešnejších a najprelomovejších.

3.1 Klasifikácia predikčných metód

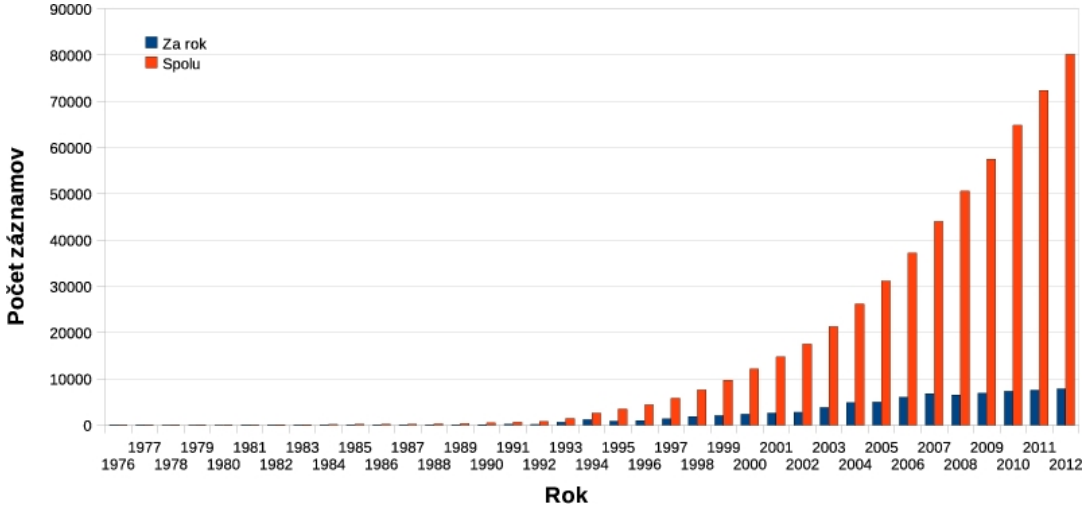
Podľa chronológie možno metódy predikcie SSP rozdeliť do 3 generácii [53] (viď tabuľka 3.1). Úspešnosť metód 1. generácie nie je vysoká, čo je dané najmä neuvažovaním globálneho kontextu preteínu ani evolučnej informácie extrahovanej z príslušnej rodiny homologických proteínových sekvencií.

Prvé metódy sa zameriavali najmä na idenifikáciu α-helixov a boli založené hlavne na modeloch popisujúcich prechody medzi helixami a coilami [22].

Neskôr sa motív sekundárnej štruktúry pre určitú aminokyselinu sa určoval na základe štatistiky, ktorá uprednostňuje ten motív, ktorý je pre danú aminokyselinu najbežnejší. Vtedajší nedostatok dát neumožňoval využiť plný potenciál štatistického prístupu. Medzi najvýznamnejšie techniky spadajúce do tejto generácie patrí metóda Chou-Fasman [12]

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

[2] Z angl. Define Secondary Structure of Proteins.



Obrázek 3.1: Graf vyjadrujúci nárast počtu záznamov v databáze PDB, dáta získané z [7].

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 Bayesovkej štatistiky a teórie informácie, ktoré sú aplikované na okno o veľkosti 17 aminokyselín (8 vľavo, 8 vpravo). Pre každú z 20 aminokyselín sa vypočítú frekvencie výskytu na danej pozícii v okne, na základe ktorých sa predikuje aminokyselina 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 [26]. GOR II pracuje s rozšírenou databázou, inak je totožná so základnou metódou GOR. Tieto metódy vo vtedajšej dobe vykazovali vyššiu úspešnosť než bola v skutočnosti kvôli zahrnutiu trénovacích sekvencií do testovacích [36].

Dalsie zlepsovaky: - celkovo, the more divergent profiele search against todays databases supposedly improves any method using alignment information by almost 4 percentage points

EVA: - EVA - automatic evaluation of automatic prediction servers - princip je nasledujuci: kazdy tyzden zober sekvencie pridane do PDB, pusti na ne vsetky dostupne metody predikcie sekundarnej struktury proteinov a vysledky porovnaj

SSpro -

PSIPRED - implementuje detailnu strategiu (of avoiding pollution of the profile through unrelated proteins), kvoli vyhnutiu sa tejto pasci musi byt datova sada urcitym spôsobom prefiltrovana ...citacia..(11)

Metódy vyvinuté v 80-tych rokoch 20. storočia možno považovať za 2. generáciu metód SSP. Vyšší výpočetný výkon dovoloval 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 metód predikcie SSP. Revolúciou v SSP bola dostupnosť rozsiahlých rodín homologickych 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 sekvenií [27], GOR V [40], ZPRED [29], či PREDATOR [21], ktorý používa metódu najbližších susedov skombinovanú s interakciou vzdialenejších aminokyselín.

Generácia	**Obdobie**	**Úspešnosť [%]**	Princíp	Mod
1.	1960 – 1980	50–55	predispozície jednotlivých aminokyselín	štatis
2.	1980 – 1990	55–62	predispozície segmentov aminokyselín	ANN,
3.	1990 – súčasnosť	70–80	využitie evolučnej informácie	ANN, HM

Tabulka 3.1: Generácie metód predikcie SSP.

Začiatkom 90-tych rokov minulého storočia začali vznikať metódy 3. generácie, ktoré sú založené na strojovom učení. Často sa používajú umelé neurónové siete, skryté Markovove modely, či klasifikátor SVM (z angl. Support Vector Machine). Klasifikátor SVM ukázal ako vhodný pre predikciu lokalizácie coilov, ktoré sú ťažko identifikovatelné štatistickými metódami [50].

Tieto metódy kombinujú silu väčších databáz a čoraz sofistikovanejších algoritmov. Základným prvkom týchto metód je využívanie evolučnej informácie – profily proteínových rodín. Bolo totiž zistené, že všetky prírodne vytvorené proteíny o dĺžke viac než 100 reziduí s viac než 35 % ..pairwise... zhodou reziduí má podobnú štruktúru. Navyše, neutrálne mutácie sú veľmi nepravdepodobné ... lebo??..., takže proteínov reálne existuje len malý zlomok. Jedným z dôsledkov je, že úseky reziduí o dĺžke povedzme 17 implicitne obsahujú dôležité informácie o globálnych interakciách, pretože profily viacnásobného zarovnania reflektujú evolučné obmedzenia [52]. PSI-BLAST a HMM ako predikčný model. Najlepšiu úspešnosť vykazujú metódy zamerané na špecifickú triedu proteínov.

Známymi predstaviteľmi sú PSIPRED [34], PHD [56], PROF [54], JPred3 [15], SSpro [51], PHDpsi [55], Copenhagen ???.

- bolo tiez zistene, ze informacia z pozicne specifickeho profilu ziskaneho zarovnanim pre konkretnu proteinovu triedu zjednodusuje objavovanie (more distant) vzdialenejsich clenov danej rodiny

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 % [66]. Nemožno nespomenúť metódu JPRED či NPS.

Podľa štúdie z roku 2009 [4], ktorá jednotnou metodológiou analyzovala úspešnost 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.

novych technik OK?

Dôležité je podotknúť, že nie je možné dosiahnuť úspešnosti blížiacej sa k 100 %. Teoretický horný limit úspešnosti je okolo 90 % [18], sčasti kvôli neistej DSSP identifikácii blízko koncov sekundárnych štruktúr, kde sa lokálne konformácie menia na základe prirodzených podmienok, no môže sa domnievať, že ide o single conformation in crystals due to packing constraints. Uvedená limitácia je taktiež spôsobená neschopnosťou predikcie sekundárnej štruktúry uvažovať terciárnu štruktúru. Sekvencia predikovaná ako helix stále môže

nadobúdať konformáciu β-sheet, ak je lokalizovaná v rámci β-sheet regiónu proteínov a jeho postranné reťazce pack well s jeho susednými reťazcami. Dramatické zmeny spôsobené vlastnou funkciou proteínu alebo prostredím, v ktorom sa nachádza, môžu taktiež meniť lokálnu sekundárnu štruktúru.

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 [2] alebo veľkosť chemických posunov [44].

3.2 Hodnotiace prístupy

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

Q_3 udáva pomer správne klasifikovaných reziduí proteínovej sekvencie do 1 z 3 tried (H, E, C) k všetkým reziduám [59]. 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 [58].

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) [57] 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 [58]. Vzhľadom k tomu, že túto mieru budem používať pri hodnotení úspešnosti SSP a jej netriviálnosti, nasledujúca časť sekcie prináša jej podrobný popis.

3.2.1 SOV

Nech s_1 a s_2 značia porovnávané segmenty sekundárnej štruktúry v konformačnom stave i (H, E 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 segment s_2 v stave i, ktorý by ich prekrýval, formálne:

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

Definícia SOV miery:

$$SOV = \sum_{i \epsilon \{H,E,C\}} SOV(i) = \frac{100}{N} \sum_{i \epsilon \{H,E,C\}} \sum_{S(i)} \left[\frac{\text{minov}(s_1, s_2) + \delta(s_1, s_2)}{\text{maxov}(s_1, s_2)} \times len(s_1) \right], \quad (3.1)$$

kde N je normalizačná hodnota:

$$N = \sum_{i \epsilon \{H,E,C\}} N(i) = \sum_{i \epsilon \{H,E,C\}} \left[\sum_{S(i)} len(s_1) + \sum_{S'(i)} len(s_1) \right], \quad (3.2)$$

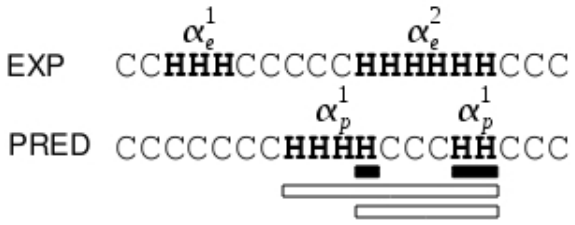
kde $len(s_1)$ vyjadruje počet reziduí v segmente s_1, $minov(s_1, s_2)$ dĺžku aktuálneho prekryvu segmentov s_1 a s_2, $maxov(s_1, s_2)$ rozsah „zjednotenia“ segmentov s_1 a s_2 a $\delta(s_1, s_2)$ je definované nasledovne:

$$\delta(s_1, s_2) = min\{maxov(s_1, s_2) - minov(s_1, s_2); minov(s_1, s_2); \lfloor len(s_1)/2 \rfloor; \lfloor len(s_2)/2 \rfloor\}, \quad (3.3)$$

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

Pre predstavu je uvedený príklad výpočtu SOV miery pre konformačný stav H pre dvojicu sekvencií zobrazenej na obrázku 3.2. Hodnota $SOV(H)$ sa na základe rovnice 3.1 vypočíta nasledovne:

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



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 vykazuje 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 [60]. Jedným z cieľov umelej inteligencie je odhaliť princíp emergentného chovania ako takého. Medzi základné prístupy blížiace sa k tomuto odhaleniu patria agentné systémy, teória chaosu, či teória celulárnych automatov.

Koncept celulárneho automatu (CA) bol vynájdený už mnohokrát pod rôznymi názvami. V matematike ide o oblasť topologickej dynamiky, v elektrotechnike sú to iteračné polia, deti ich môžu poznať ako druh počítačovej hry [61]. Modelovanie pomocou CA je principiálne jednoduché, na základe lokálneho pôsobenia ich elementov je možné vykazovať požadované globálne chovanie. V tejto kapitole bude model CA priblížený, načrtnuté historické pozadie a uvedené 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 [47]. Von Neumann sa inšpiroval prácami W. McCullocha a W.Pittsa – otcov neurónových sietí [17]. Používal 2 D CA, ktorého bunky sa mohli nachádzať v 1 z 29 stavov, okolie bolo 5-bunkové (neskôr nazývané ako *von Neumannovo*). Tento muž, ktorý pracoval aj na projekte Manhattan[1], dokázal existenciu konfigurácie zloženej z približne 200 000 buniek, ktorá sa dokáže samoreprodukovať. Takýto CA môže simulovať Turingov stroj [25]. Po von Neumannovej smrti (1957) bol jeho dôkaz zjednodušovaný.

Edgar F. Codd vytvoril jednoduchší model s 8 stavmi [14], ktorý ale neimplementoval samoreprodukčné chovanie. Tri roky po Coddovej práci, Edwin Roger Banks vytvoril elegantný 4-stavový CA v rámci svojej dizertačnej práce [5], ktorý bol schopný univerzálneho výpočtu, no opäť absentovala implementovaná samorepredukcia. John Devore vo svojej diplomovej práci významne zredukoval zložitosť Coddovho návrhu, no samoreprodukčný proces si vyžadoval príliš dlhú pásku. Až Christopher Langton upravil Coddov model do podoby schopnej vytvárať tzv. Langtonove slučky vykazujúce samoreprodukciu s minimálnym množstvom po-

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

trebných buniek, avšak za cenu absencie výpočetnej univerzality [42]. Ďalšími významnými naledovníkmi 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 20. storočia, 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 Conwaya s názvom „Game of Life“ [24]. Išlo o 2 D CA, ktorý je pomocou veľmi jednoduchých pravidiel schopný vykazovať, prenesene povedané, známky života.

Začiatkom 80-tych rokov 20. storočia 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. Podarilo sa im vytvoriť modely obyčajných diferenciálnych rovníc, akými sú napríklad rovnice prúdenia tepla, vĺn, či Navier-Stokesove rovnice prúdenia tekutín [20].

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.

4.2 Model CA

Všetky počítačové programy možno považovať v princípe za CA, pretože počítač pracuje s obmedzenou aritmetikou aj pamäťou. Väčšina CA však používa stavový priestor redukovaný na pár stavov (často len 2 stavy – 0/1) [19].

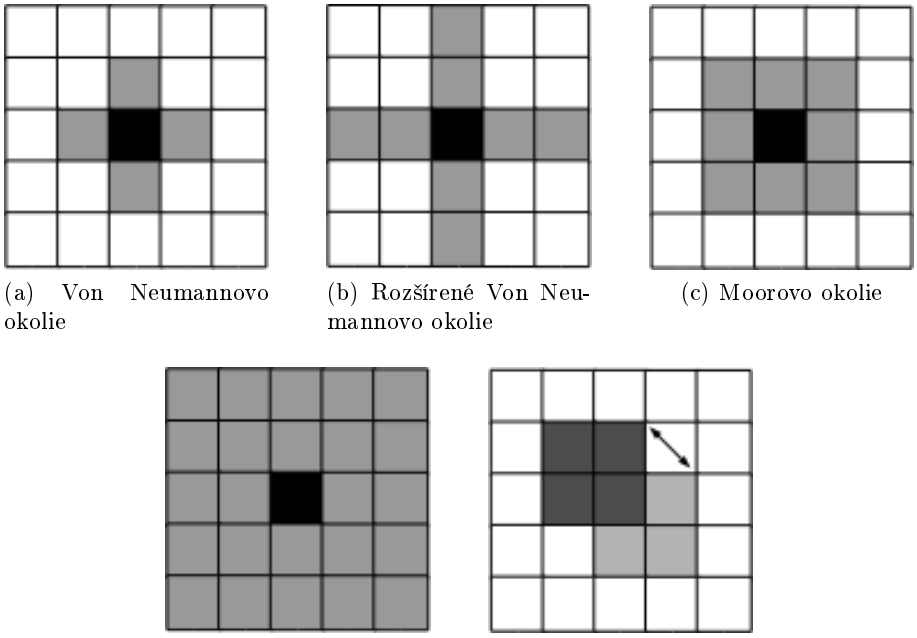
CA je dynamický systém, v ktorom je čas aj priestor diskrétny. Skladá sa z mriežky jednoduchých funkčných jednotiek – buniek, ktoré môžu nadobúdať jeden z viacerých, vopred definovaných stavov. Stavy buniek sú synchrónne aktualizované v každom kroku výpočtu CA na základe prechodovej funkcie. Formálne [23]:

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

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.

Okolie buniek CA môže byť špecifikované rôzne, obrázok 4.1 zobrazuje najčastejšie používané. Všetky vizualizované okolia sú intuitívne jasné, až na Margolusove okolie (4.1e). Tento typ okolia je špecifický tým, že plochu s bunkami je nutné rozdeliť na štvorce o veľkosti 2×2 a stav všetkých buniek v štvorci závisí len na 4 bunkách ohraničených daným štvorcom. Navyše, všetky bunky v jednotlivých štvorcových plochách majú rovnaký stav. Aby sa však nebránilo propagácii stavov buniek, celá sieť 2×2 plôch sa posunie v každom párnom kroku evolúcie automatu o jednu bunku v ose x aj y a v každom nepárnom kroku zase späť. Popísaný, pomerne zvláštny typ okolia sa úspešne využíva napríklad pri približnom riešení, už skôr spomínaných, Navier–Stokesových rovníc [62].

Stephen Wolfram, o ktorom Terry Sejnowski, odborník na neurónove siete, hovorí ako o jednom za najinteligentnejších vedcov planéty [38], definoval 4 triedy, do ktorých možno rozdeliť celulárne automaty a niektoré ďalšie výpočetné modely [64]. Príklady pravidiel



(a) Von Neumannovo okolie

(b) Rozšírené Von Neumannovo okolie

(c) Moorovo okolie

(d) Rozšírené Moorovo okolie

(e) Margolusovo okolie

Obrázek 4.1: Rôzne typy okolia CA

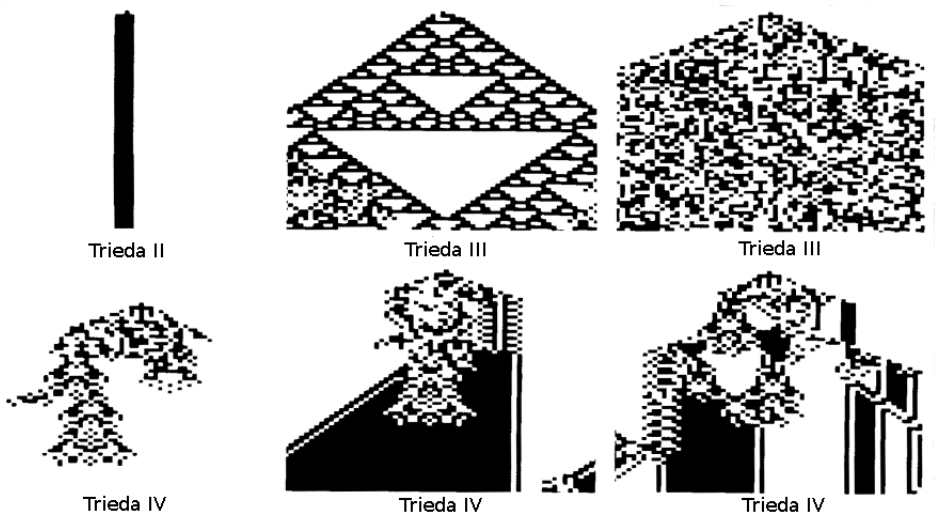
celulárnych automatov, ktoré spadajú do tried II až IV (trieda I je je triviálna), vizualizuje obrázok 4.2. Pomocou týchto tried popísal vzťah celulárnych automatov k dynamickým systémom (uvedeným v zátvorkých):

- **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).

Hlavné rysy CA sú:

- **paralelizácia** – jednotlivé stavy buniek je možné počítať paralelne
- **lokalita** – nový stav bunky závisí na aktuálnom stave bunky a savoch okolitých buniek
- **homogenita** – všetky bunky používajú rovnakú prechodovú funkciu, to však platí len pre klasické celulárne automaty (uniformné)
- **diskrétnosť** – časová aj priestorová

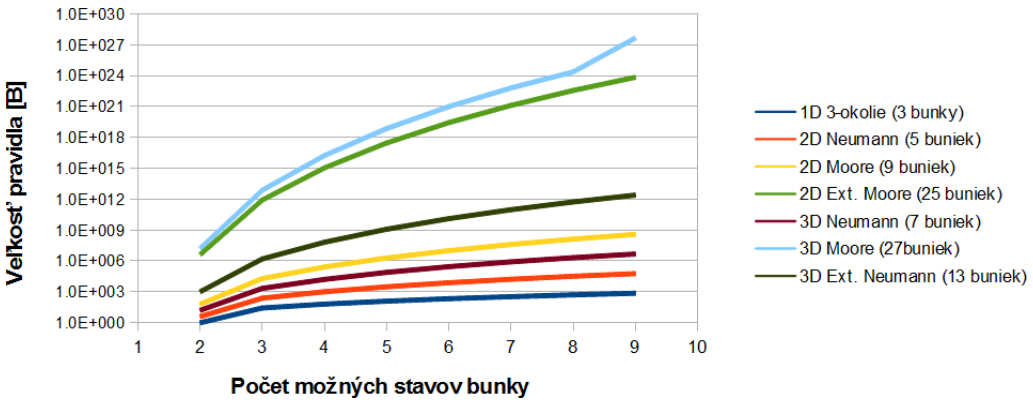
Každá navrhnutá abstrakcia reality – model, má na základe svojej špecifikácie výhody a nevýhody. Medzi svetlé stránky modelu CA patrí najmä:



Obrázek 4.2: Príklady evolúcie konfigurácie celulárnych automatov spadajúcich do tried II, III a IV.

- **jednoduchosť** – či už ide o jednoduchosť principiálnu alebo implementačnú
- **paralelizmus** – stavy buniek v nasledujúcom časovom úseku je možné počítať paralelnem čo predstavuje obrovskú výhodu a rýchlostný potenciál evolúcie CA
- **vizuálnosť** – výstupy sú väčšinou vizualizovatelné a vhodné na spracovanie príslušnými nástrojmi, napríklad ..???

Pamäťová náročnosť modelu však stúpa pri nepatrnom zväčšení okolia či zvýšení počtu stavov, ktorých môžu jednotlivé bunky nadobúdať. Porovnanie náročnosti modelu v závislosti na počte stavov a type resp. veľkosti okolia je znázornené na obrázku **??**. Diskrétnosť sa v niektorých prípadoch nehodí a môže byť kontraproduktívna. Klasický model CA má rôzne obmedzenia ako napríklad rovnaký typ okolia pre všetky bunky, tieto nedostatky riešia rôzne rozšírenia klasického konceptu CA.



Obrázek 4.3: Pamäťová náročnosť modelu celulárneho automatu v závislosti na počte stavov a type okolia.

Azda najpopulárnejším celulárnym automatom je už spomínaný celulárny automat s názvom „Game of Life“.

Je spomenutý v každej spisbe týkajúcej sa CA, pripomína vývoj spoločenstva živých organizmov. . Existuje veľké množstvo implementácii[2], kde 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 zomierajú
- živé bunky s 2 alebo 3 živými susedmi prežívajú
- živé bunky s viac než 3 živými bunkami zomierajú
- mrtvé bunky s práve 3 živými susedmi ožívajú

Paralelný výpočet buniek CA dal vznik mnohým hardwarovým riešeniam šitým na mieru špecifickým problémom, príkladom je CAM (Cellular Automata Machines) vyvinutý na MIT (Massachusetts Institute of Technology), za ktorej vývojom stoja Norman H. Margolus a Tomasso Toffoli. V dobe písania tejto práce bola aktuálna verzia CAM-8[3]

4.3 Aplikačné domény CA

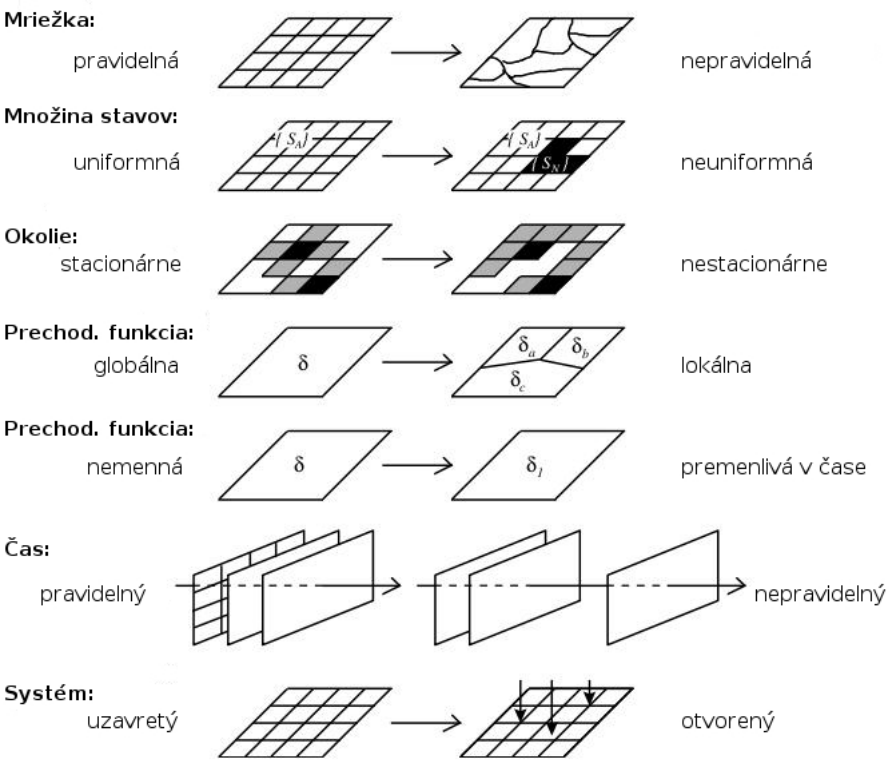
Pomerne úzka aplikačná doména celulárnych automatov sa postupom času rozrástla až do takej miery, že sa CA začali používať v každej oblasti, kde je potrebné simulovať nejaké dynamické priestorové deje, ktoré sú charakteristické svojou komplexnosťou. V praktických experimentoch sa CA využívajú v rôznych modifikáciach, málokedy sa použije model klasického celulárneho automatu. Príklady modifikácii CA sú znázornené na obrázku 4.4.

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 [65].

Model „bunkového“ automatu možno použiť aj na modelovanie biologických dráh, konkrétne na modelovanie signálnych dráh mitogénom aktivovaných proteínkynáz [39]. 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.

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

[3] Pre bližšie informácie o CAM-8 navštívte http://www.ai.mit.edu/projects/im/cam8.



Obrázek 4.4: Klasický vs. modifikovaný CA, upravené a prevzaté z [10].

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 – *O vzniku druhov prírodným výberom alebo uchovávanie prospešných plemien v boji o život*. Vrhá v ňom 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 oblasti „časovej zložitosti“. Aplikované princípy klasickej evolučnej teórie s mnohými „vylepšeniami“ hrajú významnú úlohu vo fonde vedomostí ľudstva.

Evolučné algoritmy (EA), ktoré sú postavené na myšlienkach evolučnej teórie, začali vznikať už v 50-tych rokoch 20. storočia. Výraznejší záujem však nastal až koncom 80-tych rokov minulého storočia, kedy David Goldberg významne rozšíril prácu Johna Hollanda o genetických algoritmoch (z roku 1975 [32]) v práci publikovanej v roku 1989 [28]. Značným impulzom pre popularizáciu EA bola prvá väčšia práca o genetickom programovaní, ktorej autorom je John Kozza [41].

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 EA. Pre ujasnenie pojmov je uvedený ich krátky prehľad.

Medzi základné pojmy Darwinovej evolučnej teórie patrí populácia. Populácia je množina jedincov, ktorí sú reprezentovaní svojím genetickým materiálom. V tejto práci budú voľne zamieňané pojmy genetický materiál, genóm a chromozóm, aj keď z pohľadu biológie tieto pojmy nie sú rovnocenné. Genotyp je vlastné zakódovanie genetickej informácie do určitej štruktúry. Spôsob, akým sa genotyp v danom prostredí interpretuje, ako dobre rieši nastolený problém, sa nazýva fenotyp. Jedinec s rovnakým genotypom môže mať v inom prostredí odlišnú schopnosť prežitia, inými slovami, odlišné prostredie spôsobí odlišnú interpretáciu genotypu na fenotyp. Genetický materiál sa skladá z lineárne usporiadaných génov,

v kontexte EA jeden gén kóduje jednu vlastnosť. Konkrétna vlastosť, hodnota génu, sa nazýva alela. V rámci počítačovej terminológie môžeme povedať, že každý gén reprezentuje určitý dátový typ a alely sú hodnotami daného dátového typu, génu.

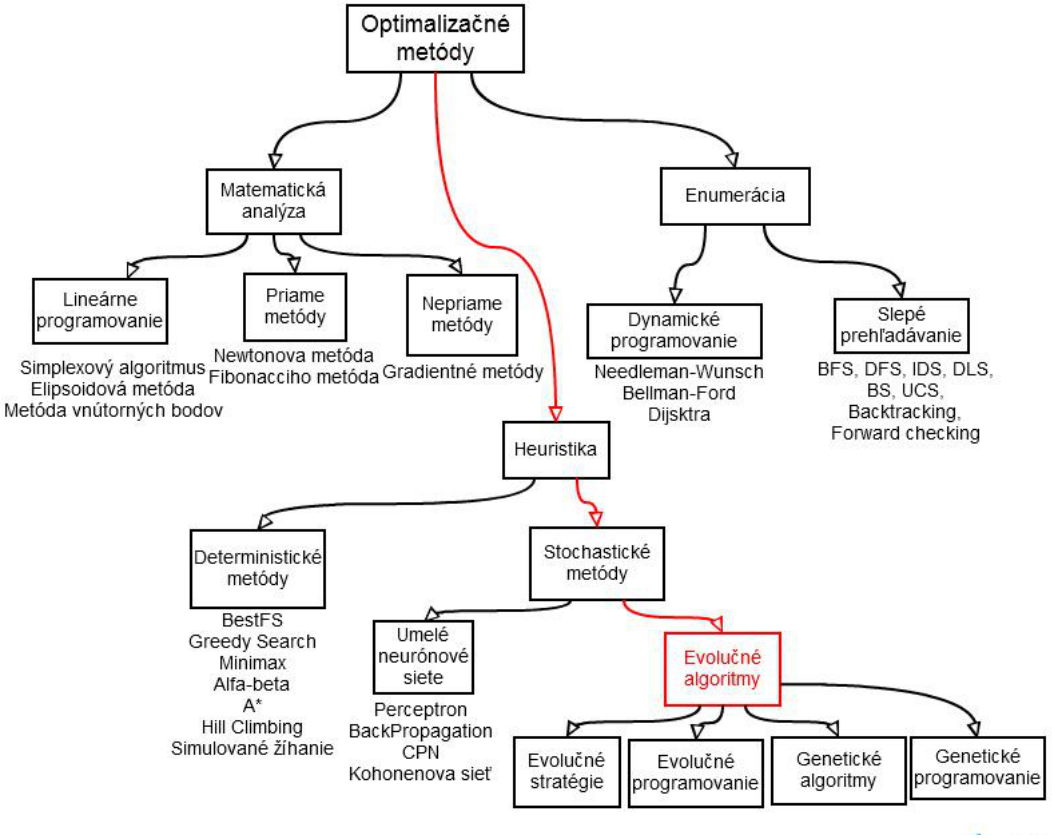
5.2 Zaradenie evolučných algoritmov

Existujú problémy, ktorých exaktný model sa nedá, nevieme alebo je veľmi náročné zostaviť, a navyše počet riešení daného problému resp. stavový priestor úlohy je obrovský. Kvôli takýmto problémom vznikol pojem *softcomputing* a všetko čo podeň spadá. Je to spôsob riešenia problémov „s rozumom“, teda nie hrubou (hard) silou ako je to prípade klasických algoritmov na prehľadávanie stavového priestoru (napr. BFS, DFS). Softcomputing rieši problémy pomocou určitej heuristiky resp. heuristickej funkcie, ktorou algoritmus „myslí“. Potrebné je dodať, že metódy softcomputingu väčšinou nenájdu najlepšie riešenie, ale len suboptimálne, ktoré však takmer vždy postačuje. Obrázok 5.1 zobrazuje zasadenie evolučných algoritmov do kontextu všetkých významných optimalizačných techník, ktorých ucelený prehľad je výborne spísaný v.

Evolučné algoritmy sú spoločným vyjadrením pre množinu moderných matematických postupov, ktoré využívajú modely evolučných procesov v prírode založených na Darwinovej evolučnej teórii popísanej na začiatku tejto kapitoly. Jednotlivé riešenia, ktoré tvoria populáciu, sa vyvíjajú na základe klasických evolučných a genetických operátorov ako je selekcia, kríženie či mutácia. EA stavajú a súčasne aj padajú na fitness funkcii, ktorá definuje schopnosť jedinca prežiť v danom prostredí resp. schopnosť riešenia riešiť daný problém. Evolučné algoritmy sú v poslednej dobe veľmi rozšírené. Prílišná popularizácia so sebou však prináša aj nerealistické očakávania. Podobne ako evolučné algoritmy, tak všeobecne aj všetky optimalizačné techniky nefungujú ako univerzálne metódy, ale každá z nich sa hodí na inú oblasť problémov a zistiť, ktorá technika je najlepšia, prípadne s akými parametrami, je už úlohou inžinirskeho návrhu.

Podstatným rozdielom oproti klasickým optimalizačným metódam je práca nie s jedným, ale s množinou riešení, na ktorú je možno aplikovať rôzne genetické a iné algoritmy. Všeobecnú schému evolučného algoritmu je možné definovať nasledovne [33]:

1. Vynuluj hodnotu počítadla generácii $t = 0$.
2. Náhodne vygeneruj počiatočnú populáciu $P(0)$.
3. Vypočítaj ohodnotenie ($fitness$) každého indivídua v počiatočnej populácii $P(0)$.
4. Vyber dvojice indivíduí z populácie $P(t)$ a vytvor ich potomkov $P'(t)$.
5. Vytvor novú populáciu $P(t+1)$ z pôvodnej populácie $P(t)$ a množiny potomkov $P'(t)$.
6. Zväčši hodnotu počítadla generácii o jedna ($t := t + 1$).
7. Vypočítaj ohodnotenie ($fitness$) každého indivídua v populácii $P(t)$.
8. Ak je t rovné maximálnemu počtu generácii alebo je splnené iné ukončovacie kritérium, vráť ako výsledok populáciu $P(t)$; inak pokračuj krokom číslo 4.



Obrázek 5.1: Klasifikácia evolučných algoritmov v kontexte optimalizačných techník. Soft-computing spadá pod stochastické heuristické metódy.

5.3 Evolučné stratégie

Keď v roku 1963 začali Hans-Paul Schwefel a Ingo Rechenberg na Technickej univerzite v Berlíne s napodobňovaním vývoja v prírode, boli presvedčení, že ich metóda najlepšie aproximuje evolúciu v živej prírode. Preto svoju metódu nazvali, celkom všeobecne, evolučné stratégie (ES). Postupom času sa však ukázalo, že tento spôsob rieši len určitý typ úloh, hlavne v stavebnom a strojnom inžinierstve. Genetické algoritmy nie sú teda podradené evolučným stratégiam, ako sa domnievali, ale naopak, svojou popularitou ich zatieňujú [37]. Genóm v rámci ES je zložený z génov, ktoré sú reprezentované reálnymi číslami, z čoho vyplýva implementačná diferenciácia genetických operátorov. Mutačný operátor je väčšinou implementovaný pomocou pripočítania hodnoty podľa Gaussovej funkcie. Takýto spôsob mutácie rieši jeden problém genetických algoritmov – malé zmeny genotypu nemusia viesť k malým zmenám fenotypu[1].

5.4 Genetické operátory

Evolučné procesy z biológie boli aplikované a svojím spôsobom interpretované v teórii evolučných algoritmov. Ide o pekný príklad medzioborového transféru informácií. V evolučnom

Trošku rozviesť

[1] Tento problém sa dá obísť napríklad pomocou Grayovho kódovania

procese je teda nutná značná variácia genómu. Genetickými operátormi sú selekcia (prirodzený výber) a rekombinačné operátory – kríženie (obrázok 5.2) a mutácia (obrázok 5.3).

5.4.1 Selekcia

Selekcia je základným evolučným operátorom, určuje, ktoré riešenie v populácii riešení prežije, a ktoré nie. Reprezentuje prirodzený výber popísaný Darwinom. Rozoznávame 3 najpoužívanejšie typy selekcie a ich varianty:

- koleso šťastia – jednotlivým jedincom sa priradí pravdepodobnosť výberu do ďalšej generácie na základe hodnoty fitness funkcie, „lepší“ jedinci budú vyberaný do ďalšej genrácie s vyššou pravdepodobnosťou než „horší“ jedinci
- turnaj – je založený na náhodnom výbere n-tíc jedincov a ich súboji, v ktorom sú zbraňami hodnoty fitness funkcie, víťaz je vo väčšine variant tejto selekcie vybraný do ďalšej generácie vždy, no môže byť vyberaný s pravdepodobnosťou menšou než 1
- „najlepší vyhráva“ – je najjednoduchším typom selekcie, v každej generácii preferuje len tých najstatnejších jedincov, jedincov umiestnených na čelných priečkach rebríčka zostavovaného komisiou, ktorá hodnotí statnosť jedinca na základe hodnoty fitness funkcie. Popisovaný spôsob výberu je vhodný v prípade, keď fitness funkcia nemá veľa extrémov, pretože evolučné algoritmy založené na tomto type selekcie nevedia riešiť tzv. klamné problémy a multimodálne funkcie, pretože v populácii nezachovávame diverzitu, vyberáme jedincov s najlepšou hodnotou fitness funkcie. Evolučné algoritmy založené na tomto type selekcie častkrát predbežne konvergujú, teda uviaznu v lokálnom extréme.

1

Pri reálnych praktických aplikáciach je len málokedy možné sa stretnúť s čistým jedným typom selekcie, takmer vždy sa používajú ich možné variácie a kombinácie. Pre priblíženie, existujú aj prístupy, ktoré pracujú s dvoma populáciami kvôli zachovaniu rôznorodosti populácie a dochádza k migrácii medzi populáciami. Každá populácia je však založená na inej fitness funkcii [48].

So selekciou úzko súvisí obnova populácie. Po vyhodnotení hodnôt fitness funkcie jedincov populácie a selekcii jedincov, ktorí „prežijú“, máme viac možností ako nahradiť aktuálnu populáciu novou. Rozlišujeme 2 základné prístupy:

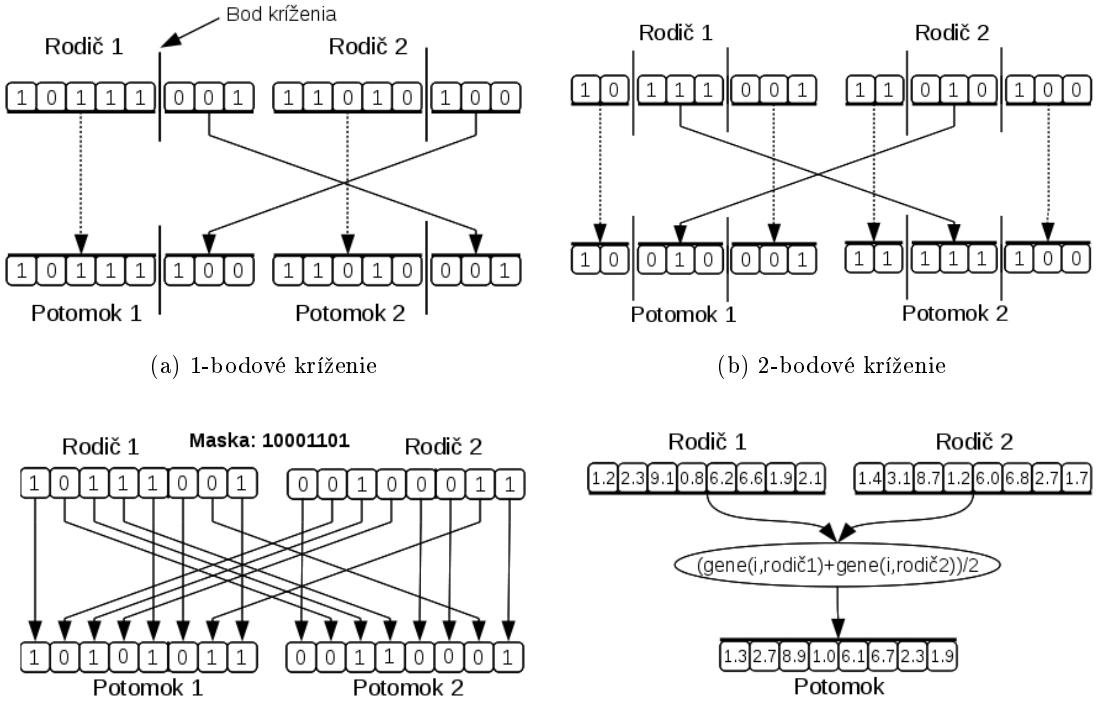
- **úplná obnova populácie** – dochádza k vymieraniu rodičov, teda celá generácia je nahradená novou
- **čiastočná obnova populácie** (steady state) – potomkami sa nahradi len určitá časť jedincov

5.4.2 Kríženie

Kríženie je základným rekombinačným operátorom, pomocou ktorého sa mieša genetická informácia 2 jedincov. K tomuto operátoru možno pristupovať viacerými spôsobmi, ktorých vizuálna podoba je zobrazená na obrázku 5.2:

- **n-bodové kríženie** – najpoužívanejší typ kríženia, väčšinou sa používa kríženie jednobodové alebo dvojbodové, závisí samozrejme na danom probléme a veľkosti chromozómu

- **kríženie maskou** – náhodne sa vygeneruje bitová maska, ktorej dĺžka je zhodná s dĺžkou chromozómu a noví jedinci sa tvoria tak, že gény na pozíciach obsahujúcich „0“ zdedí prvý potomok a gény na pozíciach obsahujúcich „1“ zdedí potomok druhý
- **aritmetické kríženie** – sa využíva najmä pri evolučných stratégiach, kde sú gény reprezentované reálnymi číslami, noví jedinci sa tvoria na základe aplikácie nejakého aritmetického operátoru (väčšinou priemer) na gény rodičov



(a) 1-bodové kríženie

(b) 2-bodové kríženie

(c) Kríženie maskou

(d) Aritmetické kríženie (pomocou priemeru)

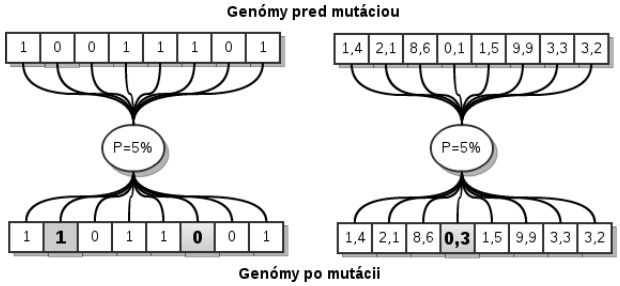
Obrázek 5.2: Rôzne typy evolučného operátoru kríženia, prevzaté z [9].

5.4.3 Mutácia

Mutácia je motorom evolučných algoritmov, je založená na náhodnej zmene génu chromozómu jedinca, mutácia zabezpečuje vznik nových vlastností, zmenu génov. Rozlišujeme v zásade 2 typy mutácie (ilustruje obrázok 5.3):

permutačné chromozómy

- **inverzia génu** – vhodné a použitelné len pri binárnej reprezentácii génov, teda génoch, ktoré môžu existovať len v 2 navzájom rôznych alelách
- **pripočítanie hodnoty rozloženia pravdepodobnosti** – využíva sa pri reprezentácii génov reálnymi číslami, k hodnote génu sa pripočíta hodnota daná určitým rozdelením pravdepodobnosti



Obrázek 5.3: Mutácia genómu, ľavá časť obrázku reprezentuje aplikáciu operátoru mutácie pomocou inverzie génu, pravá časť pomocou pripočítania hodnoty rozloženia pravdepodobnosti.

Kapitola 6

Návrh predikčného systému

nazvy kapitol

Navrhnutý systém z veľkej časti vychádza z dvoch prác venujúcich sa SSP s využitím CA [11] [63]. Modifikované sú tie časti systému, ktoré majú potenciál zlepšiť úspešnosť predikcie SSP. Ide hlavne o spôsob klasifikácie jednotlivých reziduí do jednej z troch motívov sekundárnej štruktúry proteínov.

trochu rozviest

Predikčným modelom je celulárny automat, ktorého prechodová funkcia bola suboptimálne parametrizovaná pomocou evolučných algoritmov, konkrétne evolučných stratégii.

Boli navrhnuté 2 prechodové funkcie. Prvá je zhodná s prechodovou funkciu vytvorenou Choprom a spol., no bola implementovaná najmä kvôli porovnávaniu v experimentálnej fáze. Druhou je metóda využívajúca jednak klasické Chou-Fasmanove koeficienty, druhak tzv. konformačné koeficienty. Ide tiež o štatistický popis pravdepodobnosti výskytu určitej aminokyseliny v určitom konformačnom stave resp. motíve sekundárnej štruktúry.

Významnou je experimentálna časť práce, ktorá sa zameriava na:

- experimentálne zistenie optimálnej veľkosti okolia, ktoré bude využívať prechodová funkcia celulárneho automatu, podľa práce pánov .. .. by mal byť limit veľkosti okolia 17 (8 vľavo, 8 vpravo), kedy sa úspešnosť signifikantne nezvyšuje ...cite...
-

Aby sa predikčný model mal o čo oprieť, definujeme štatistické vlastnosti reziduí.

6.1 Štatistický popis reziduí

Štatistika je 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 [13] 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. [67] definujeme konformačnú triedu pre všetky aminokyseliny a všetky konformačné stavy (H, E, C). Nech $P = p_1, p_2, \ldots, p_n$ je primárna štruktúra proteínu (sekvencia aminokyselín) a $S = s_1, s_2, \ldots, s_n$ 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:

[1] Z angl. Structure Transition.

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

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):

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

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 Použitý celulárny automat

Sekvenciu aminokyselín proteínov reprezentuje 1 D CA, ktorého bunky modelujú jednotlivé reziduá aminokyselín. Bunky môžu nadobúdať 1 z 3 stavov (H, E, C). Veľkosť okolia nie je pevná, evolučný algoritmus zisťuje optimálny rádius. Štatistické 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.

6.2.1 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:

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

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

$$R^i_{t+1,j} = P^i_{t+1,j} + \alpha \cdot CC^i_j \tag{6.12}$$

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

6.3 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

Implementácia systému

7.0.1 Implementačné prostriedky

7.0.2 Konfigurácia systému

7.0.3 Obmedzenia systému

7.0.4 Webové rozhranie

Webové rozhranie je minimalistické, no spĺňa účel. Zobrazené je na obr. .., kde sú 2 panely, ktoré etc...

Kapitola 8

Korektor PSIPREDu

Primárnou snahou experimentovania s navrhnutým modelom bolo zlepšiť úspešnosť jedného v súčasnosti najlepších nástrojov, PSIPRED-u. Dôležitá je vhodná parametrizácia modelu, pre ktorej dosiahnutie bola najskôr vykonaná optimalizácia parametra veľkosti okolia, ktoré uvažuje prechodová funkcia CA. Následne bol zistený optimálny maximálny počtet krokov CA, ktoré môžu pri evolúcii pravidla CA jednotlivé riešenia nadobúdať. Po nájdení najvhodnejších hodnôt týchto 2 parametrov boli vykonané experimenty predikujíce sekundárnu štruktúru proteínu v spolupráci s nástrojom PSIPRED. Uvažované boli 2 varianty:

1. Primárna predikcia pomocou navrhnutého systému CASSP a následná oprava nie príliš vierohodných predikcii pomocou nástroja PSIPRED.

2. Primárna predikcia pomocou nástroja PSIPRED a následná oprava nie príliš vierohodných predikcií pomocou navrhnutého systému CASSP.

Implementované sú 2 spôsoby opravy predikcie primárneho prediktora:

1. Použitie predikcie sekundárneho prediktora pre reziduum, ktorého vierohodnosť predikcie je nižšia než zadaný prah.

2. Použitie predikcie sekundárneho prediktora pre celú proteínovú sekvenciu, ak priemerná vierohodnosť reziduí v rámci opravovanej sekvencie je nižšia než zadaný prah.

V oboch prípadoch je dôležité správne stanoviť prah opravy primárnej predikcie pomocou predikcie sekundárnej. Pre zistenie vhodného prahu bola vykonaná jeho optimalizácia pre obe varianty.

Optimálne parametrizované model bol natrénovaný a výsledky porovnané s vybranými existujúcimi nástrojmi.

Umožnený prístup k výpočetným a úložným zariadeniam vedeckých skupín prispievajúcich do Národnej sieťovej infraštruktúry MetaCentrum, vytvorenej v rámci programu „Rozsiahla infraštruktúra pre projekty výskumu, vývoja a inovácii“ (LM2010005), je vysoko cenený.

prepracovať podakovanie

8.1 Trénovacie a testovacie dátové sady

Ak chceme porovnávať výkonnosť modelu s inými modelmi, je nutné, aby úspešnosti vštkých porovnávaných modelov boli počítané na základe rovnakej testovacej dátovej sady.

V oblasti predikcie sekundárnej štruktúry proteínových sekvencií sú najpoužívanejšími dátovými sadami používanými ako na trénovanie, tak aj na testovanie, RS126 a CB513.

Uvedené experimenty pracovali s 3 dátovými sadami.

RS126 [56] - ...

Porovnávanie úspešnotí Dátová sada RS126 bola prvý krát použitá v článku Burkharda Rosta a Chrisa Sandera z roku 1993 [56]. Podľa ich slov ide o nehomológnu dátovú sadu, nehomológnosť definovali tak, že žiadne 2 proteiny v dátovej sade nesmú mať viac než 25 % zhody v sekvenciách pri ich dĺžke viac než 80 reziduí. Nevýhodou (okrem malého počtu sekvencií) je že dátová sada RS15 obsahuje páry proteínových sekvencií, ktoré sú nepochybne podobné pri provnávaní sofistikovanejšími metódami než obyčajnou percentouálnou zhodou. Navyše, výpočet percentuálnej zhody je závislý ako na dĺžke zarovnania, tak aj na zložení sekvencií. Takže, 2 sekvencie podobného, ale nezvyčajného aminokyselinového zloženia, môžu mať vysokú percentuálnu zhodu, aj keď sú nehomológne/unrelated [16].

cb513

Dátovú sada CB513 uviedli páni Barton a Cuff vo svojej štúdii z roku 1999 [16].

Našťastie, existujú techniky, ktoré nemajú nevýhody obyčajnej percentuálnej zhody. Príkladom môže byť metóda, ktorá najskôr zarovná sekvencie štandardným algoritmom dynamického programovania (napríklad algoritmus Needleman-Wunsch) a pre toto zarovnanie sa získa skóre.

Poradie aminokyselín v každej proteínovej sekvencii je randomizovaný a následne je vykonané zarovnanie pomocou spomínaného dynamického programovania takejto pomiešanej sekvencie. Tento proces je opakovaný typicky aspoň 100 krát a priemer a smerodatná odchýlka jednotlivých skóre. is calculated.

SD alebo Z skóre porovnania je dané nasledovne: $(V - \bar{x})/\sigma$.

Z počiatočnej dátovej sady sa odstránili multisegmentové domény, odstránené boli aj sekvencie, ktorých štruktúry získané pomocou röntgenovej kryštalografie nemali dostatočné rozlíšenie (2,5 Å).

Následne boli vyhodené sekvencie podobné s nejakou sekvenciou z dátovej sady RS126 pri podobnosti $SD >= 5$.

Sekvencie, ktoré nemali úplnú DSSP definíciu boli taktiež odstránené. Následne bola prefiltrovaná dátová sada spojená s RS126. Miera podobnosti nie je schopná zachytiť všetky homológne sekvencie, na ďalšie porovnávanie sekvencií bol použitý algoritmus SCOP [45]. Výsledkom je dátová sada CP513.

Chou-Fasmanove koeficienty a konformačné koeficienty pre nejednoznačné aminokyseliny (B,Z,X) boli nahradené priemernými hodnotami, pre B je to priemer priemer hodnôt aminokyseliny asparagín a kyseliny asparagovej, pre Z priemer hodnôt aminokyseliny glutamín a kyseliny glutámovej, pre J priemer hodnôt aminokyseliny leucín a izoleucín.

pdb_vyber

Treťou použitou dátovou sadou je celkom rozsiahly subset približne 5300 proteínových sekvencií. Zoznam proteínových sekvencií bol vytvorený .... [30].

Získanie sekvencií pomocou webovej služby MRS [31]. V niektorych sekvenciach nahradene o za C, u za C (). Vzhľadom k rozsiahlosti dátovej sady bola v kontexte tejto práce využívaná iba na testovanie.

warum??

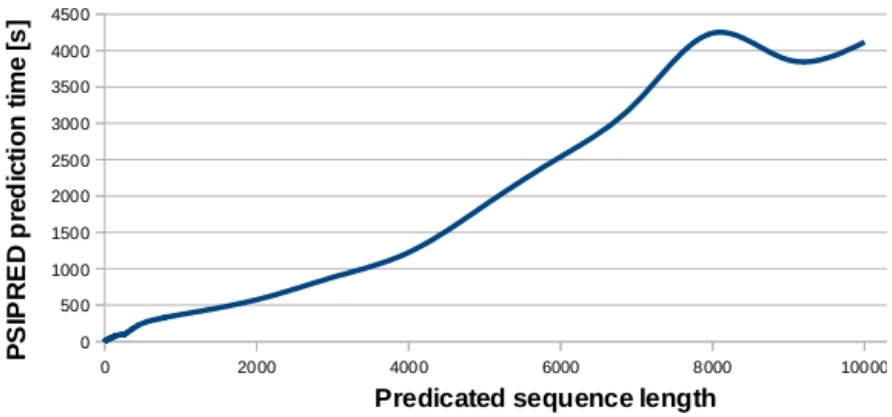
PDBselect je zoznam reprezentatívnych proteínovych sekvencií s nizkou mutúal/základnou sekvenčnou podobnosťou/identitiou

PDBselect (http://bioinfo.tg.fh-giessen.de/pdb select/) is a list of representative protein chains with low mutual sequence identity selected from the protein data bank (PDB) to

enable unbiased statistics. The list increased from 155 chains in 1992 to more than 4500 chains in 2009. PDBfilter-select is an

Keďže vývoj v oblasti bioinformatiky je veľmi rýchly a počet záznamom v PDB rastie exponenciálne, možno považovať databázy RS126 a CB513 za zastaralé a pre reálne praktické použitie by sa siahlo po novšej, viac aktualizovanej databáze. No pre základnú charakteristiku navrhnutého modelu sú tieto 2 dátové sady dostačujúce, navyše, takmer všetky nástroje predikcie sekundárnej štruktúry spomenuté v kapitole 3 pracujú práve s týmito dátovými sadami, takže je možné priame porovnanie úspešnosti.

Pri trénovaní modelu, ktorý predikuje sekundárnu štruktúru proteínu v spolupráci s nástrojom PSIPRED, a vzhľadom k pomerne vysokej časovej zložitosti tohto nástroja (viď obrázok 8.1), boli PSIPRED predikcie všetkách sekvencií dátových sád predpočítané.



Obrázek 8.1: meeeh

8.2 Okrajové podmienky a inicializácia modelu

okrajove podmienky salanda

Pri modelovaní javov pomocou celulárneho automatu je dôležitá definícia okrajových podmienok - podmienok, ktoré popisujú situácie, kedy okolie aktuálnej bunky nie je kompletné, teda na okrajov automatu. Je vhodné definovať bunky mimo celulárneho automatu pomocou novovytvorených aminokyselín s optimálnymi vlastnosťami z pohľadu úspešnosti predikcie. Práca [63] sa určením takejto aminokyseliny zaoberala a prišla s výsledkom optimálnej aminokyseliny X300, ktorej vlastnosti sú nasledovné:

napisat presnu rovnicu gaussovej funkcie + asi aj obrazok pre ilustraciu

inicial. modelu - na zaklade Gaussovej funkcie - normalizovanej

Na inicializácii jedincov/chromozómov v rámci evolučného algoritmu v podstate nezáleží, no pre rýchlosť konvergencie je dôležité sa zaoberať aj touto časťou systému. Je intuitívne jasné, že vplyv reziduí v tesnom okolí predikovanej aminokyseliny bude vyšší než vplyv reziduí vzdialenejších. Platí to najmä pri motívoch alpha helixu, no beta sheety asto vznikajú globálnou interakciou, ktorú by sa mal pokúsiť emulovať navrhnutá model celulárneho automatu.

Navrhnutá je inivializácia váh/vplyvu jednotlivých okolitých reziduí na základe normalizovanej Gaussovej funkcie pre strednú hodnotu $\mu = 0$ a smerodatnú odchýlku $\sigma = 0.399$ (hodnota $f(0) \doteq 1$):

$$\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{8.1}$$

8.3 Metodika vyhodnocovania úspešnosti

Pri korektnom popise prediktora je veľmi dôežitý spôsob vyhodnocovania jeho úspešnosti.

Azda najdôležitejšou podmienkou je, aby trénovacie a testovacie dáta nekorelovali, (unbiased), čo je ťažké dosiahnuť v reálnych podmienkach, kedy dát často nie je dostatok a ich rozloženie je neznáme.

Na mieste je teda otázka, ako ideálne rozdeliť dátovú sadu na trénovaciu a testovaciu tak, aby sme získali čo možno najdôveryhodnejšiu hodnotu úspešnosti?

Simone Borra a Agostino Di Ciaccio vo svojej práci [8] došli k záveru, že 10-fold cross-validácia pre vyšší počet vzorkov než 100, vykazuje najvernejšiu hodnotu chyby prediktoru pre reálne dátové sady. It is known that K-fold CV is biased estimate of Err and that by increasing the number of folds we can reduce the bias. Leave-One-Out cross-validácia (LOO) teoreticky vykazuje lepší/unbiased výsledok, no pri testovaní vzhľadom k tomu, že testovacia dátová sada obsahuje iba 1 prvok, sa prejavuje veľká variabilita, čo robí problémy pri selekcii najlepšieho ďielčieho modelu. 1O-fold cross-validácia bude teda použitá pre vyhodnocovanie jednotlivých modelov. Cross-validácia bude spustená x-krát a spriemerovaná pre získanie ešte hodnovernejšej hodnoty.

8.4 Optimálna veľkosť okolia CA

.. experiment zisťujúci optimálnu veľkosť okolia celulárneho automatu

8.5 Maximálny počet krokov CA

8.6 Predikcia v spoluprácii s nástrojom PSIPRED

blah...

8.6.1 PSIPRED ako hlavný prediktor

optimálna hodnota prahu...

8.6.2 PSIPRED ako sekundárny prediktor

optimálna hodnota prahu

Kapitola 9

Záver

citacie datasetov: [56], [16], [30]

cross-validation stuff: [8]

Proteíny sú základnými stavebnými kameňmi života na Zemi, starajú sa 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] Protein's Four Levels. [online],[cit. 2013-03-20].
URL
http://eleventeengreen.wordpress.com/2012/11/19/proteins-four-levels/

[2] A., L.; A., M. S.: Addition of contact number information can improve protein secondary structure prediction by neural networks. *EXCLI Journal*, ročník 08, 2009: s. 66–73.

[3] 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 978-80-902906-2-0, 740 s.

[4] Bagos, P. G.; Tsaousis, G. N.; Hamodrakas, S. J.: How Many 3D Structures Do We Need to Train a Predictor? *Genomics, Proteomics & Bioinformatics*, ročník 7, č. 3, 2009: s. 128 – 137, ISSN 1672-0229.

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

[6] Berman, H. M.; Westbrook, J.; Feng, Z.; aj.: The Protein Data Bank. *Nucleic Acids Res*, ročník 28, 2000: s. 235–242.

[7] Berman, H. M.; Westbrook, J.; Feng, Z.; aj.: RCSB PDB Statistics. 2012, [online],[cit. 2013-01-05].
URL http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html

[8] Borra, S.; Ciaccio, A. D.: Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis*, ročník 54, č. 12, 2010: s. 2976–2989, ISSN 0167-9473.

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

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

[11] 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.

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

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

[14] Codd, E. F.: *Cellular Automata*. Orlando, FL, USA: Academic Press, Inc., 1968, ISBN 978-0-1217-8850-4.

[15] Cole, C.; Barber, J. D.; Barton, G. J.: The Jpred 3 secondary structure prediction server. *Nucleic Acids Research*, ročník 36, č. 2, 2008: s. 197–201.

[16] Cuff, J. A.; Barton, G. J.: Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, ročník 34, č. 4, 1999: s. 508–519, ISSN 1097-0134.

[17] Delorme, M.; Mazoyer, J.: *Cellular Automata: a parallel model*, ročník 460. Springer, 1998, ISBN 978-0-7923-5493-2.

[18] Dor, O.; Zhou, Y.: Achieving 80secondary structure prediction by large-scale training. *Proteins: Structure, Function, and Bioinformatics*, ročník 66, č. 4, 2007: s. 838–845, ISSN 1097-0134, doi:10.1002/prot.21298.
URL http://dx.doi.org/10.1002/prot.21298

[19] Ermentrout, B. G.; Edelstein-Keshet, L.: Cellular Automata Approaches to Biological Modeling. *Journal of Theoretical Biology*, ročník 160, č. 1, jan 1993: s. 97–133, ISSN 0022-5193.

[20] Fredkin, E.; Toffoli, T.: Collision-based computing. kapitola Conservative logic, London, UK, UK: Springer-Verlag, 2002, ISBN 978-1-85233-540-8, s. 47–81.

[21] Frishman, D.; Argos, P.: Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein engineering*, ročník 9, č. 2, 1996: s. 133–142.

[22] Froimowitz, M.; Fasman, G. D.: Prediction of the Secondary Structure of Proteins Using the Helix-Coil Transition Theory. *Macromolecules*, ročník 7, č. 5, 1974: s. 583–589.

[23] 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.

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

[25] Gardner, M.: *Wheels, life, and other mathematical amusements*. Freeman, 1983, ISBN 978-0-7167-1589-4.

[26] Garnier, J.; Gibrat, J. F.; Robson, B.: GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol*, ročník 266, 1996: s. 540–553.

[27] 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.

[28] Goldberg, D. E.: *Genetic Algorithms in Search, Optimization and Machine Learning*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., první vydání, 1989, ISBN 978-0-2011-5767-5.

[29] Granseth, E.; Viklund, H.; Elofsson, A.: ZPRED: Predicting the distance to the membrane center for residues in alpha-helical membrane proteins. In *ISMB (Supplement of Bioinformatics)*, 2006, s. 191–196.

[30] Griep, S.; Hobohm, U.: PDBselect 1992-2009 and PDBfilter-select. *Nucleic Acids Research*, ročník 38, č. Database-Issue, 2010: s. 318–319.

[31] Hekkelman, M. L.; G., V.: MRS: a fast and compact retrieval system for biological data. *Nucleic Acids Res*, ročník 33, č. Web Server issue: s. W766–9+.

[32] Holland, J. H.: *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI, USA: University of Michigan Press, 1975.

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

[34] Jones, D. T.: Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *Journal of Molecular Biology*, ročník 292, 1999: s. 195–202.

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

[36] Kabsch, W.; Sander, C.: How good are predictions of protein secondary structure? *FEBS Letters*, ročník 155, č. 2, 1983: s. 179 – 182, ISSN 0014-5793.

[37] Kalátová, E.; Dobiáš, J.: Evoluční algoritmy. [online],[cit. 2011-03-11].
URL http://www.kiv.zcu.cz/studies/predmety/uir/gen_alg2/E_alg.htm

[38] Kapoun, J.: Nový druh vědy si dobře rozumí s byznysem. 2005, [online],[cit. 2013-01-07].
URL http://scienceworld.cz/technologie/novy-druh-vedy-si-dobe-rozumi-s-byznysem-1715

[39] 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.

[40] Kloczkowski, A.; Ting, K.; Jernigan, R.; aj.: Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins*, ročník 49, č. 2, 2002: s. 154–66.

[41] Koza, J. R.: *Genetic Programming: On the Programming of Computers by Means of Natural Selection (Complex Adaptive Systems)*. The MIT Press, první vydání, December 1992, ISBN 978-0-262-11170-5.

[42] Langton, C. G.: Self-reproduction in cellular automata. *Physica D: Nonlinear Phenomena*, ročník 10, č. 1–2, 1984: s. 135–144, ISSN 0167-2789.

[43] Liu, X.; Fan, K.; Wang, W.: The number of protein folds and their distribution over families in nature. *Proteins: Structure, Function, and Bioinformatics*, ročník 54, č. 3, 2004: s. 491–499, ISSN 1097-0134.

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

[45] Murzin, A. G.; Brenner, S. E.; Hubbard, T.; aj.: SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, ročník 247, č. 4, 1995: s. 536–540, ISSN 0022-2836.

[46] Nečas, O.: *Obecná biologie pro lékařské fakulty*. H & H Vyšehradská, 2000, ISBN 978-80-86022-46-8.

[47] von Neumann, J.: *Theory of Self-Reproducing Automata*, ročník 160. Illinois: University of Illinois Press, 1966, ISBN 978-0-598-37798-0.

[48] Park, T.; Ryu, K. R.: A Dual-Population Genetic Algorithm for Adaptive Diversity Control. *Evolutionary Computation, IEEE Transactions on*, ročník 14, č. 6, december 2010: s. 865–884, ISSN 1089-778X.

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

[50] Pham, T. H.; Satou, K.; Ho, T. B.: Support Vector Machines for prediction and analysis of beta and gamma-turns in proteins. *Journal of Bioinformatics and Computational Biology*, ročník 03, č. 02, 2005: s. 343–358.

[51] Pollastri, G.; Przybylski, D.; Rost, B.; aj.: Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function, and Bioinformatics*, ročník 47, č. 2, 2002: s. 228–235, ISSN 1097-0134.

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

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

[54] Rost, B.; Eyrich, V. A.: EVA: Large-Scale Analysis of Secondary Structure Prediction. ročník 5, 2001: s. 192–199.

[55] Rost, B.; Sander, C.: Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, ročník 90, č. 16, 1993: s. 7558–7562, ISSN 0027-8424.

[56] Rost, B.; Sander, C.: Prediction of protein secondary structure at better than 70584–599.

[57] 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.

[58] 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.

[59] 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.

[60] Sipper, M.: *Evolution of Parallel Cellular Machines: The Cellular Programming Approach*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2001, ISBN 978-3-540-62613-8.
URL http://www.cs.bgu.ac.il/~sipper/papabs/epcm.pdf

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

[62] Tyler, T.: The Margolus neighbourhood. [online],[cit. 2013-01-07].
URL http://cell-auto.com/neighbourhood/margolus/

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

[64] Wolfram, S.: *A New Kind of Science*. Wolfram Media, January 2002, ISBN 978-1-57955-008-8, 1197 s.

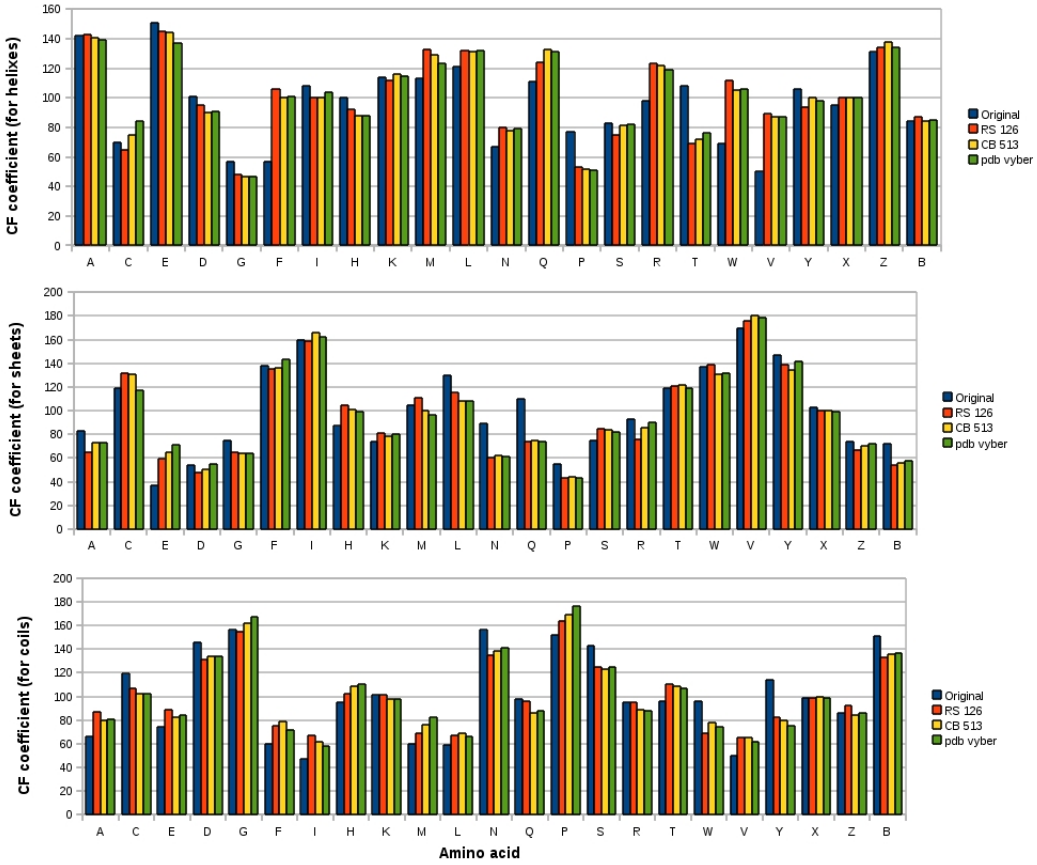
[65] Xiao, X.; Shao, S.; Ding, Y.; aj.: Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids*, ročník 30, 2006: s. 49–54, ISSN 0939-4451.

[66] 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.

[67] 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, nov 2005: s. 2346–2352, ISSN 0167-8655.

Príloha A

Štatistické vlastnosti dátových sád



Obrázek A.1: meeeh

Príloha B

Výsledky experimentov