

# 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 jako doplňkový nástroj ke stávajícím metodám predikce sekundární struktury proteinů.

## 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 proteinů, 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  
20. března 2013

## Poděkování

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

© Vladimír Brigant, 2013.

*Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů.*

# Obsah

<b>1</b>	<b>Úvod</b>	<b>2</b>
<b>2</b>	<b>Proteíny</b>	<b>3</b>
2.1	Tvorba proteínov . . . . .	3
2.2	Štruktúra proteínov . . . . .	3
2.3	Funkcie proteínov . . . . .	5
2.4	Významné projekty súvisiace s analýzou proteínov . . . . .	5
<b>3</b>	<b>Predikcia sekundárnej štruktúry proteínov</b>	<b>6</b>
3.1	Klasifikácia predikčných metód . . . . .	6
3.2	Hodnotiace prístupy . . . . .	8
<b>4</b>	<b>Celulárne automaty</b>	<b>10</b>
4.1	Historické pozadie . . . . .	10
4.2	Model CA . . . . .	11
4.3	Významné CA . . . . .	12
4.4	Aplikácie modelu CA . . . . .	13
<b>5</b>	<b>Evolučné algoritmy</b>	<b>14</b>
5.1	Biologické pojmy v EA kontexte . . . . .	14
5.2	Zaradenie evolučných algoritmov . . . . .	15
5.3	Genetické operátory . . . . .	16
<b>6</b>	<b>Návrh predikčného systému</b>	<b>17</b>
6.1	Štatistický popis reziduí . . . . .	17
6.1.1	Chou-Fasmanove parametre . . . . .	17
6.1.2	Konformačné triedy . . . . .	18
6.2	Použitý celulárny automat . . . . .	19
6.2.1	Prechodová funkcia . . . . .	19
6.3	Použitý evolučný algoritmus . . . . .	19
<b>7</b>	<b>Záver</b>	<b>21</b>

# 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 [28]) a experimentálne metódy sú časovo aj finančne náročné, čo podnietilo vznik predikčných metód štruktúry proteínov. Medzikrok 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.

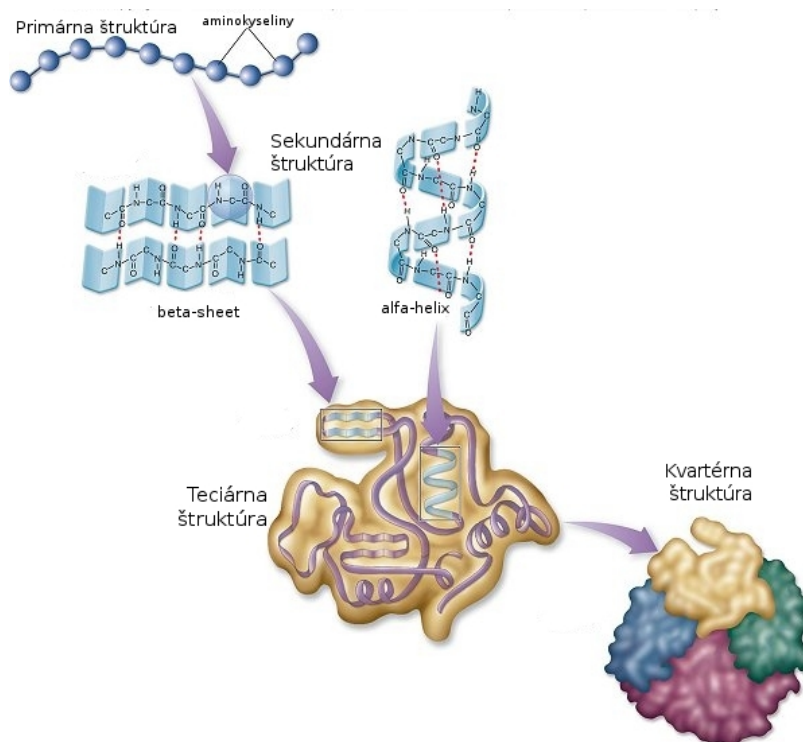
Najväčším problémom CA je vhodné určenie prechodovej funkcie. Nie je možné „odskúšať“ všetky a určiť najlepšiu, preto prichádzajú na scénu optimalizačné techniky, v našom prípade genetický algoritmus, ktorý pri hľadaní suboptimálneho riešenia používa princípy evolučného výberu a genetiky (viac v kapitole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

Vlastným návrhom predikčného systému, použitím štatistických a experimentálnych charakteristík aminokyselín, celulárneho automatu a evolučne inšpirovaných techník, sa zaoberá kapitola 6. Záver (kapitola 7) obsahuje krátke zhrnutie práce.

3

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 odvodzovaná podľa kódujúcej sekvencie nukleotidov DNA (sekvenovanie DNA je metodicky jednoduchšie [29]).
2. **Sekundárna štruktúra** – časti polypeptidového reťazca, ktoré tvoria 2 základné štruktúrne pravidelnosti –  $\alpha$ -helix (H) a  $\beta$ -sheet (E). Reziduá aminokyselín, ktoré nepatria ani do jedného motívu, sa označujú z historických dôvodov ako Coil (C). Ako  $\alpha$ -helix označujeme takú konformáciu, kedy reťazec vytvára skrutkovicové usporiadanie, v  $\beta$ -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 zložitý, 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.



Obrázok 2.2: Úrovne štruktúry proteínov, prevzaté z [1].



## 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 [29]. 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 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<sup>1</sup>). 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<sup>2</sup>, 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 [31].

---

<sup>1</sup>Z angl. Human Genome Project, domovská stránka projektu: [http://www.ornl.gov/sci/techresources/Human\\_Genome/home.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml).

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

## Kapitola 3

# Predikcia sekundárnej štruktúry proteínov

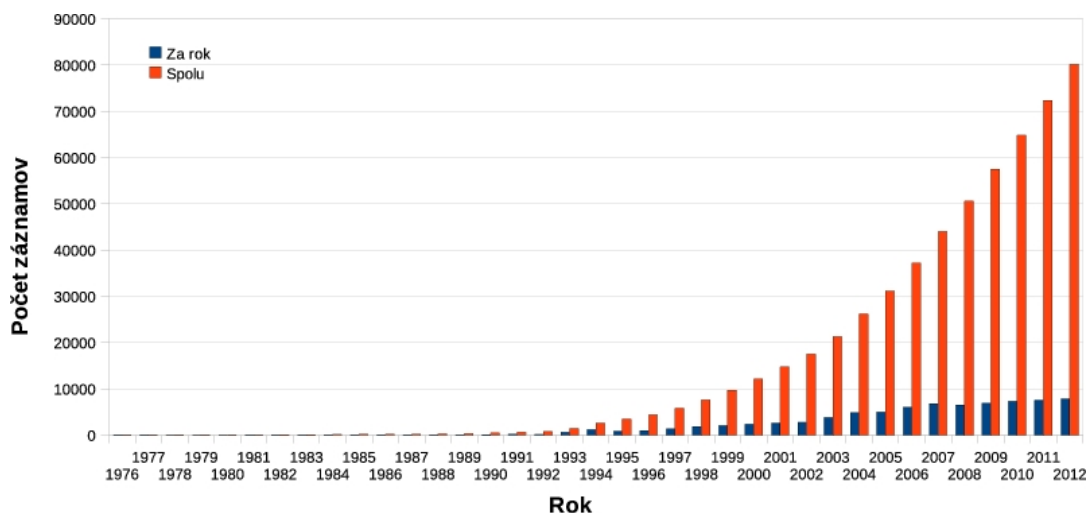
Držiteľ dvoch Nobelových cien (1954, 1962), Linus Pauling, bol prvý, kto predpovedal motívy sekundárnej štruktúry proteínov (SSP) [32]. Koncom 50-tych rokov bola po prvý krát experimentálne zistená štruktúra proteínu (pomocou röntgenovej kryštalografie<sup>1</sup>). Rozkvet experimentálneho zisťovania 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 zavedenej? v [5]. V súčasnosti je počet experimentálne zistených záznamov štruktúry proteínov približne 80 000, existuje teda databáza, 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 [23]. SSP problém znie: majme proteínovú sekvenciu s aminokyselinami  $\{S_1, S_2, \dots, S_n\}$ ; urči pre každú  $S_i$  motív sekundárnej štruktúry –  $\alpha$ -helix (H),  $\beta$ -sheet (E) alebo Coil (C). 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 generácií [33] (viď tabuľka 3.1). Úspešnosť metód 1. generácie nie je vysoká, čo je dané najmä neuvažovaním globálneho kontextu proteínu ani evolučnej informácie extrahovanej z príslušnej rodiny homologických proteínových sekvencií. Motív sekundárnej štruktúry pre určitú aminokyselinu sa určoval na základe štatistiky, ktorá uprednostňuje ten motív, ktorý je pre danú aminokyselinu najbežnejší. Tieto metódy trpeli nedostatkom dát. Medzi najvýznamnejšie metódy patrí Chou-Fasman [10] a GOR. 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 teórie informácie, ktoré sú aplikované na okno o veľkosti 17 aminokyselín (8 vľavo, 8 vpravo). Pre každú z 20 aminokyselín sa vypočíta frekvencia na danej pozícii v okne, na základe ktorých sa

<sup>1</sup>Metóda zisťovania polohy jednotlivých atómov molekúl za pomoci röntgenového žiarenia, ktoré nám dovoľuje „vidieť“ rádovo v jednotkách nanometrov.



Obrázek 3.1: Graf vyjadrujúci nárast počtu záznamov v databáze PDB, prevzaté z [6].

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 [19]. GOR II je základná metóda GOR pracujúca s rozšírenou databázou. Tieto metódy v tej dobe vykazovali vyššiu úspešnosť než bola v skutočnosti kvôli zahrnutiu tréningových sekvencií do testovacích [24].

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

Na prelom 2. a 3. generácie metód predikcie SSP možno zaradiť algoritmy rozšírené o ďalšie informácie o aminokyselinách, napr. tvar, veľkosť alebo fyzikálno-chemické vlastnosti. Patrí sem napríklad metóda najbližších susedov, kde sekundárna štruktúra sa určí na základe štruktúry najpodobnejších sekvení [20], GOR V, ZPRED, či PREDATOR, ktorý používa metódu najbližších susedov skombinovanú s interakciou vzdialenejších aminokyselín.

Generácia	Obdobie	Úspešnosť [%]
1.	1960 – 1980	50–55
2.	1980 – 1990	55–62
3.	1990 – ?	70–80

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 či klasifikátor SVM (z angl. Support Vector Machine). Známymi predstaviteľmi sú PSIPRED, PHD a PROF.

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

Podľa štúdie z roku 2009 [3], analýzy úspešnosti algoritmov predikcie SSP, existujúce algoritmičné techniky nemôžu byť naďalej vylepšované iba pridávaním nehomologických sekvencií do tréningovej dátovej sady, tzn. mali by sme zamerať na používanie nových techník pri vývoji prediktorov SSP s lepšou úspešnosťou.

## 3.2 Hodnotiace prístupy

Dôležitým prvkom pri vývoji metód predikcie SSP sú postupy merajúce úspešnosť hodnotenej metódy. Medzi najpoužívanjšie 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 [36]. 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 [35].

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) [34] 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 [35]. 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, 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 prekrývajúci sa segment  $s_2$  v stave  $i$ , formálne:

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

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

SOV miera pre konformačný stav  $i$ :

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

kde  $N$  je normalizačná hodnota:

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

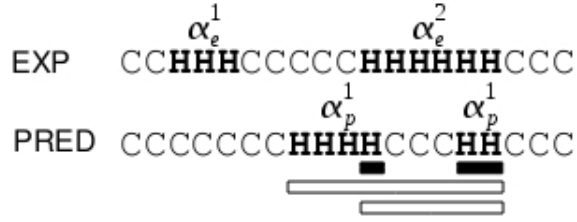
kde  $\text{len}(s_1)$  vyjadruje počet reziduí v segmente  $s_1$ ,  $\minov(s_1, s_2)$  dĺžku aktuálneho prekryvu segmentov  $s_1$  a  $s_2$ ,  $\maxov(s_1, s_2)$  rozsah „zjednotenia“ segmentov  $s_1$  a  $s_2$  a  $\delta(s_1, s_2)$  je definované nasledovne:

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

$$\lfloor \text{len}(s_1)/2 \rfloor, \lfloor \text{len}(s_2)/2 \rfloor\}, \quad (3.6)$$

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

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



Obrázek 3.2: Ilustrácia výpočtu  $\text{SOV}(\text{H})$ . 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 simulácií sa stalo základným nástrojom pri odkrývaní tajov našej Zeme. V prírode sa stretávame s rôznymi príkladmi chovania, ktoré vykazujú 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, sieťnicu alebo imunitný systém [37]. 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 [38]. 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 aplikáčnych 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 [30]. Von Neumann sa inšpiroval prácami W. McCullocha a W. Pittsa – otcov neurónových sietí [13]. Používal 2D 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<sup>1</sup>, dokázal existenciu konfigurácie zloženú z približne 200 000 buniek, ktorá sa dokáže samoreprodukovať tzn. takýto CA môže simulovať Turingov stroj [18]. Po von Neumannovej smrti v roku 1957 bol jeho dôkaz zjednodušovaný, Za zmienku stojí Coddov celulárny automat s 8 stavmi [12]. Najelegantnejší dôkaz pomocou 4 stavov však priniesol v roku 1971 vo svojej dizertačnej práci Edwin Roger Banks [4]. Následovníkmi von Neumanna v štúdiu 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

---

<sup>1</sup> Krycí názov pre utajený americký vývoj atómovej bomby počas 2. svetovej vojny.

Conwaya s názvom „Game of Life“ [17]. Išlo o 2D 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, vln, či Navier-Stokesove rovnice prúdenia tekutín [15].

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 Norman H. Margolus a Tomasso Toffoli.

## 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 na pár stavov redukovaný stavový priestor (často len 2 stavy – 0/1) [14].

CA je dynamický systém, v ktorom je čas aj priestor diskretný. Skladá sa z mriežky jednoduchých funkčných jednotiek – buniek, ktoré môžu nadobúdať jeden z viacerých, vopred definovaných stavov. Stavy buniek sú synchronne aktualizované v každom kroku výpočtu CA na základe prechodovej funkcie. Formálne [16]:

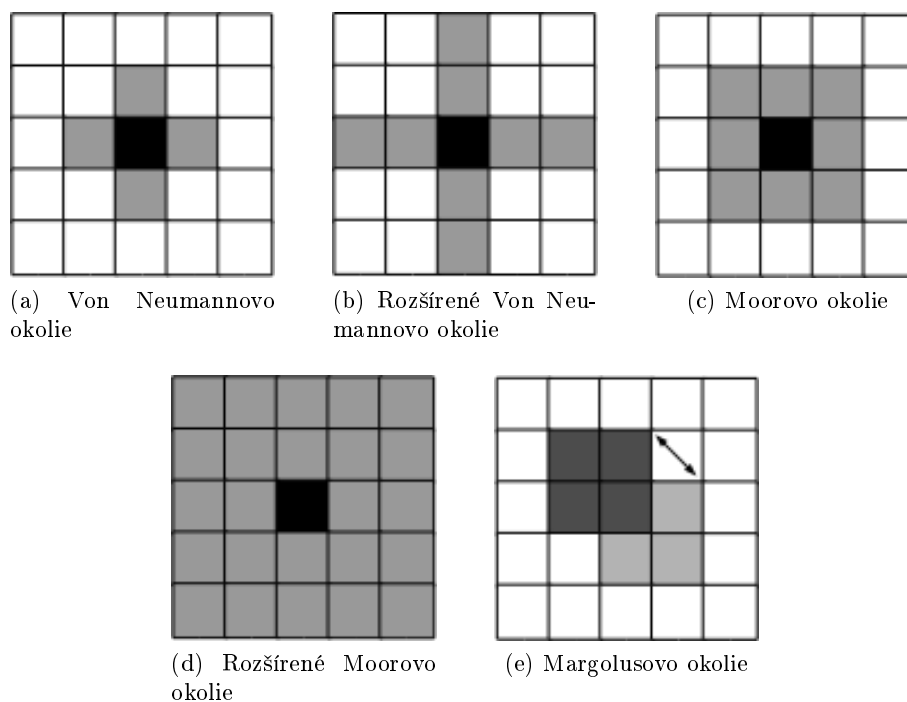
$$s^{(t+1)} = f(s^{(t)}, s_N^{(t)}), \forall i, j, \quad (4.1)$$

kde  $s^{(t+1)}$  reprezentuje stav bunky danej pozíciou  $i$  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  $2 \times 2$  a stav všetkých buniek v štvorci závisí len na 4 bunkách ohraničených daným štvorcem. Navyše, všetky bunky v jednotlivých štvorcových plochách majú rovnaký stav. Aby sa však nebránilo propagácii stavov buniek, celá sieť  $2 \times 2$  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 [39].

Stephen Wolfram, o ktorom Terry Sejnowski, odborník na neurónove siete, hovorí ako o jednom z najinteligentnejších vedcov planéty [25], definoval 4 triedy [41], do ktorých možno rozdeliť celulárne automaty a niektoré ďalšie výpočetné modely. Pomocou týchto tried popísal vzťah celulárnych automatov k dynamickým systémom (uvedený v zátvorkách):

- **Trieda I** – počiatočné konfigurácie evolvujú do stabilného, homogénneho stavu. Akákoľvek náhodnosť počiatočnej konfigurácie mizne (limitné body).



Obrázek 4.1: Rôzne typy okolia CA

- **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).

### 4.3 Významné CA

Je spomenutý v každej spisbe týkajúcej sa CA, pripomína vývoj spoločenstva živých organizmov. Ide o už spomínaný celulárny automat Johna Conwaya, „Game of Life“. Existuje veľké množstvo implementácií<sup>2</sup>. 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

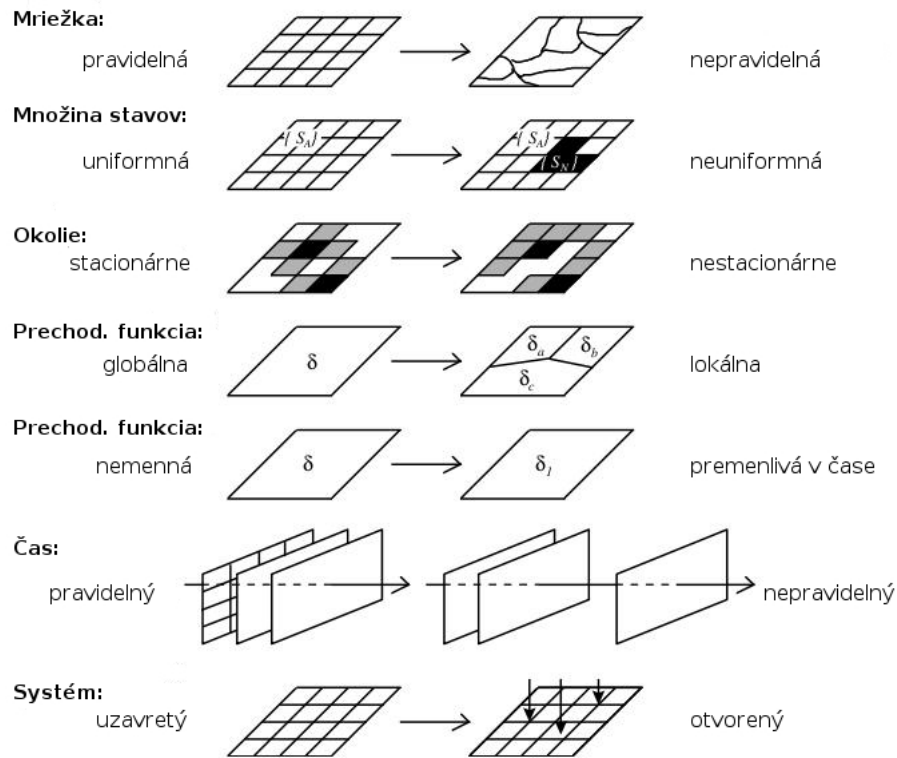
<sup>2</sup> Interaktívnu implementáciu v podobe appletu možno nájsť na adrese: <http://www.bitstorm.org/gameoflife/>.



Medzi ďalšie významné CA patrí napríklad Coddov automat, hardwarový CAM, či CA vykazujúci Langtonove Q-slučky.

## 4.4 Aplikácie modelu 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 modifikáciách, málokedy sa použije model klasického celulárneho automatu. Príklady modifikácií CA sú znázornené na obrázku 4.2.



Obrázek 4.2: Klasický vs. modifikovaný CA, upravené a prevzaté z [8].

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 [42].

Model „bunkového“ automatu možno použiť aj na modelovanie biologických dráh, konkrétne na modelovanie signálnych dráh mitogénom aktivovaných proteínkynáz [26]. 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.

## 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 ucelený pohľad na vývoj druhov oslobodený od spirituality a náboženských predstáv svojej a predchádzajúcich dôb a hlavou plnou prírodovedných informácií 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 rodiať 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 koncom 80-tych rokov minulého storočia, kedy David Goldberg významne rozšíril prácu Johna Holland z roku 1975 o genetických algoritmoch [22], v práci publikovanej v roku 1989 [21]. Značným impulzom pre popularizáciu EA bola prvá väčšia práca o genetickom programovaní, ktorej autorom je John Koza [27].

### 5.1 Biologické pojmy v EA kontexte

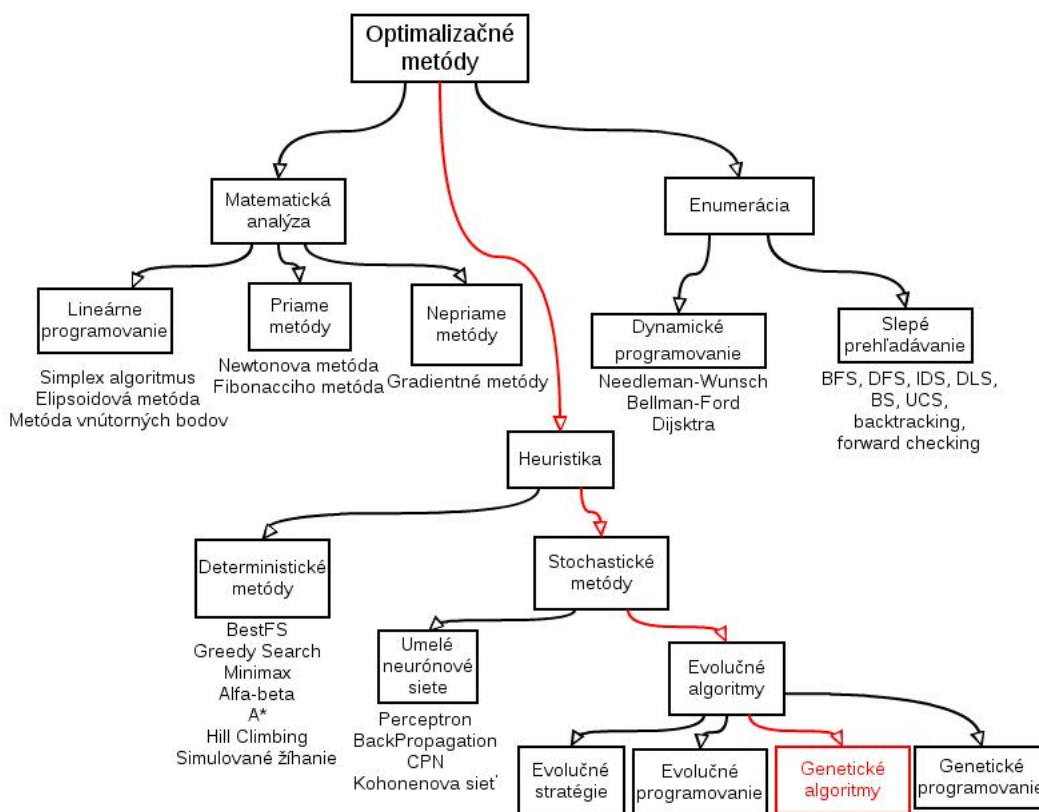
Vo zvyšku tejto práce sa budú vyskytovať pojmy pochádzajúce z biológie, no sémantika nie všetkých je zhodná so sémantikou v kontexte evolučných algoritmov. Pre ujasnenie pojmov je uvedený ich krátky prehľad.

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 kon-

texte EA jeden gén kóduje jednu vlastnosť. Konkrétna vlastnosť, 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

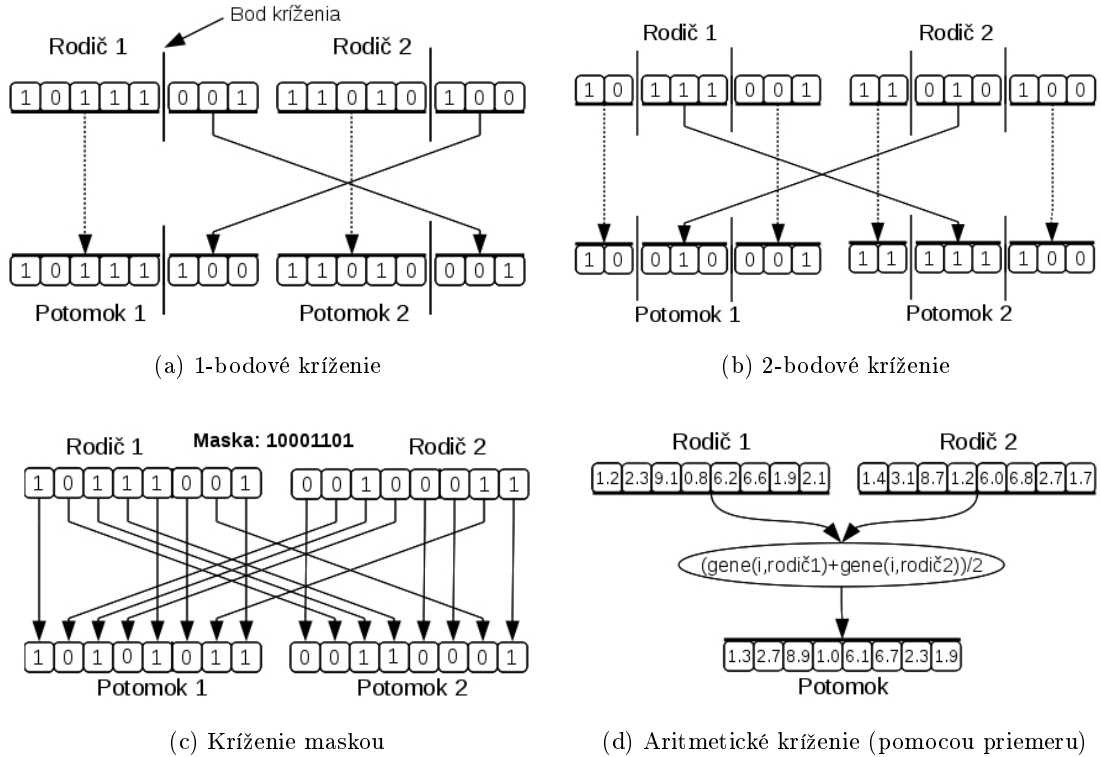
Existujú problémy, ktorých exaktný model sa nedá, nevieme alebo je veľmi náročné zostaviť, a navyše stavový priestor resp. počet riešení daného problému je obrovský. Kvôli takýmto problémom vznikol pojem *softcomputing* a všetko čo podneň spadá. Je to spôsob riešenia problémov „s rozumom“, teda nie hrubou (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, resp. genetických algoritmov do kontextu všetkých významných optimalizačných techník.



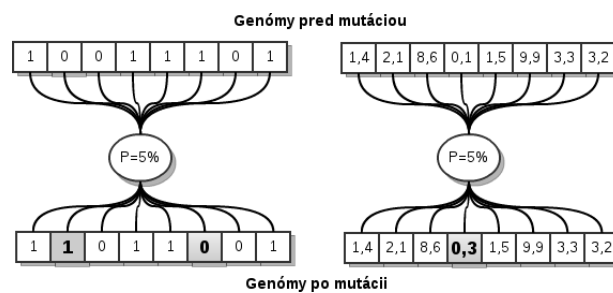
Obrázok 5.1: Klasifikácia genetických algoritmov v kontexte optimalizačných techník. Soft-computing spadá medzi stochastické heuristické metódy.

### 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 (obrázok 5.2) a mutácia (obrázok 5.3).



Obrázok 5.2: Rôzne typy evolučného operátora kríženia, prevzaté z [7].



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

## Kapitola 6

# Návrh predikčného systému

Navrhnutý systém z veľkej časti vychádza z 2 prác venujúcich sa SSP pomocou CA [9] [40]. 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 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 [11] je definovaný tzv. parameter konformačnej preferencie aminokyseliny  $j$  ku konformačnému stavu  $i$ , Chou-Fasmanov parameter  $P_j^i$ :

$$P_j^i = \frac{f_j^i}{\langle f_j^i \rangle}, \quad (6.1)$$

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.

$$f_j^i = \frac{n_j^i}{n^i} \quad (6.2)$$

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

$$\langle f_j^i \rangle = \frac{\sum_{\forall k \in AK} f_k^i}{n_j} \quad (6.3)$$

### 6.1.2 Konformačné triedy

Na základe práce Guang-Zheng Zhanga a spol. [44] definujeme konformačnú triedu pre všetky aminokyseliny a všetky konformačné stavy (H, E, C). Nech  $P = p_1, p_2, \dots, p_n$  je primárna štruktúra proteínu (sekvencia aminokyselín) a  $S = s_1, s_2, \dots, s_n$  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:

$$N_{\alpha\text{-ukončenie}} = N_{ST_{HB}} + N_{ST_{HC}} \quad (6.4)$$

$$N_{\alpha\text{-začiatok}} = N_{ST_{BH}} + N_{ST_{CH}} \quad (6.5)$$

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  $CP_{j,\alpha\text{-ukončenie}}$ ,  $CP_{j,\alpha\text{-začiatok}}$ ,  $CP_{j,\beta\text{-ukončenie}}$ ,  $CP_{j,\beta\text{-začiatok}}$ ,  $CP_{j,\text{Coil-ukončenie}}$  a  $CP_{j,\text{Coil-začiatok}}$ . Výpočet  $CP_{j,\alpha\text{-ukončenie}}$  (ostatné konformačné preferencie sa získajú analogicky):

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

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

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

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

$$P_j = \frac{N_j}{N} \quad (6.8)$$

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

---

<sup>1</sup>Z angl. Structure Transition.

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

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 1D 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:

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

$P_{t+1,j}^i$  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:

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

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, E, C):

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

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

## 6.3 Použitý evolučný algoritmus

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

$$C = [s, \alpha, r, w_{-r}, w_{-r+1}, \dots, w_{r-1}, w_r] \quad (6.13)$$

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



## Kapitola 7

### Záver

Proteíny sú základnými stavebnými kameňmi života na Zemi, starajú sa o podstatnú časť biologických funkcií 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] 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.
- [3] 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.
- [4] Banks, E. R.: Information processing and transmission in cellular automata. Technická zpráva, Cambridge, MA, USA, 1971.
- [5] Berman, H. M.; Westbrook, J.; Feng, Z.; aj.: The Protein Data Bank. *Nucleic Acids Res*, ročník 28, 2000: s. 235–242.
- [6] 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](http://www.rcsb.org/pdb/static.do?p=general_information/pdb_statistics/index.html)
- [7] Brigant, V.: *Evoluční návrh simulátoru založeného na celulárních automatech*. Bakalářská práce, FIT VUT v Brně, Brno, 2011.
- [8] Cecchini, A.; Rinaldi, E.: The multi - cellular automaton: a tool to build more sophisticated models. A theoretical foundation and a practical implementation. 1999.
- [9] 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.
- [10] Chou, P. Y.; ; Fasman, G. D.: Prediction of protein conformation. *Biochemistry*, ročník 13, č. 2, Leden 1974: s. 222–245.
- [11] Chou, P. Y.; Fasman, G. D.: Conformational parameters for amino acids in helical,  $\hat{I}_c$ -sheet, and random coil regions calculated from proteins. *Biochemistry*, ročník 13, č. 2, Leden 1974: s. 211–222.
- [12] Codd, E. F.: *Cellular Automata*. Orlando, FL, USA: Academic Press, Inc., 1968, ISBN 978-0-1217-8850-4.

- [13] Delorme, M.; Mazoyer, J.: *Cellular Automata: a parallel model*, ročník 460. Springer, 1998.
- [14] 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 0022-5193.
- [15] 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.
- [16] 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.
- [17] Gardner, M.: Mathematical Games The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, ročník 223, 1970: s. 120–123.
- [18] Gardner, M.: *Wheels, life, and other mathematical amusements*. Freeman, 1983.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] Holland, J. H.: *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI, USA: University of Michigan Press, 1975.
- [23] 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.
- [24] 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.
- [25] 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>
- [26] 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.
- [27] 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.

- [28] 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.
- [29] Nečas, O.: *Obecná biologie pro lékařské fakulty*. H & H Vyšehradská, 2000, ISBN 978-80-86022-46-8.
- [30] von Neumann, J.: *Theory of Self-Reproducing Automata*, ročník 160. Illinois: University of Illinois Press, 1966.
- [31] Pennisi, E.: Genomics. ENCODE project writes eulogy for junk DNA. *Science*, ročník 337, č. 6099, 2012: s. 1159, 1161, ISSN 1095-9203.
- [32] Rost, B.: Review: Protein Secondary Structure Prediction Continues to Rise. *J. Struct. Biol.*, ročník 134, 2001: s. 204–218.
- [33] Rost, B.: Protein Prediction - Part 1: Structure. University Lecture, 2011, [online],[cit. 2012-12-05].
- [34] 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.
- [35] 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.
- [36] 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.
- [37] 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>
- [38] Toffoli, T.; Margolus, N.: *Cellular automata Machines, A New Environment For Modeling*. Cambridge, MA: MIT Press, 1987.
- [39] Tyler, T.: The Margolus neighbourhood. [online],[cit. 2013-01-07].  
URL <http://cell-auto.com/neighbourhood/margolus/>
- [40] Šalanda, V.: *Predikce sekundární struktury proteinu pomocí celulárního automatu*. Bakalářská práce, FIT VUT v Brně, Brno, 2012.
- [41] Wolfram, S.: *A New Kind of Science*. Wolfram Media, January 2002, ISBN 978-1-57955-008-8, 1197 s.
- [42] 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.
- [43] 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.

- [44] 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.