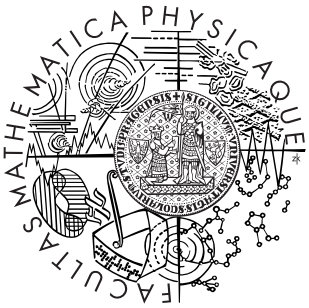
Univerzita Karlova v Praze
Matematicko - fyzikální fakulta

DIPLOMOVÁ PRÁCE



Katarína Figurová

Úlohy globální optimalizace v praxi

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Doc. RNDr. Petr Lachout, CSc.

Studijní program: Matematika

Studijní plán: Pravděpodobnost, matematická statistika a ekonometrie

Poďakovanie

Ďakujem vedúcemu dipolomovej práce Doc. RNDr. Petrovi Lachoutovi, CSc. za jeho trpezlivosť a za cenné pripomienky a návrhy pri písaní tejto práce. Rada by som poďakovala aj Mgr. Jozefovi Varju za osvetlenie fyzikálnej stránky problému, za jeho nadšenie a trpezlivosť. V neposlednom rade ďakujem celej rodine a priateľom za láskavú podporu a ich neustály optimizmus.

Čestné prehlásenie

Prehlasujem, že som svoju diplomovú prácu napísala samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičiavaním práce.

V Prahe dňa 5.8.2007 Katarína Figurová

vlastnoruční podpis

Obsah

Abstrakt

Název práce: Úlohy globální optimalizace v praxi
Autor: Katarína Figurová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Doc. RNDr. Petr Lachout, CSc.
e-mail vedoucího: lachout@karlin.mff.cuni.cz
Abstrakt: Mnoho úloh z rôznych odvetví môžeme previesť na problém globálnej optimalizácie. V tejto práci je sformulovaná úloha nájdenia globálneho extrému, uvedený stručný prehľad a rozdelenie metód globálnej optimalizácie a rovnako aj ich využitie v niektorých praktických problémoch. Predpoveď proteínovej trojrozmernej štruktúry na základe znalosti lineárnej sekvencie aminokyselin je jednou zo známych aplikácii. Tento problém je ekvivalentný problému nájdenia minima energetickej funkcie. Po uvedení základných informácií o proteínoch, funkcii potenciálnej energie, a použitých modeloch, nasleduje popis niekoľkých algoritmov použitých pri predpovedi proteínovej štruktúry.
Klíčová slova: globálna optimalizácia, problém skladania proteínov

Abstract

Title: Global optimization in practice
Author: Katarína Figurová
Department: Department of Probability and Mathematical Statistics
Supervisor: Doc. RNDr. Petr Lachout, CSc.
Supervisor's e-mail address: lachout@karlin.mff.cuni.cz
Abstract: Many practical applications can be formulated as global optimization problems. In this work the global optimization problem will be reviewed and some of the methods that have been proposed. Predicting the 3D form of protein from their linear sequence is one of the applications of global optimization. This problem can by formulated as a problem of finding a global minimum of energy function. Some basic information about proteins, their potential energy function and used models is given. Next, we described some of the algorithms used for determining the structure of proteins.
Keywords: global optimization, protein folding problem

Kapitola 1

Úvod

Množstvo úloh z praxe vedie k nájdeniu globálneho extrému - najlepšej možnosti spomedzi niekoľkých alternatív. Cieľom tejto práce je poukázať práve na použiteľnosť matematických metód globálnej optimalizácie pri riešení praktických úloh z rôznych oblastí.

Jednou zo zaujímavých aplikácií je príklad simulácie skladania proteínov, ktorý vedie na úlohu minimalizácie energetickej funkcie systému. Konkrétne určeniu proteínovej troj-dimenzionálnej štruktúry, na základe znalosti jeho postupnosti aminokyselín, je v tejto práci venovaná hlavná pozornosť. Náš záujem je smerovaný hlavne k použitým optimalizačným metódam.

Problém skladania proteínov je veľmi komplexný, preto je dôležité nezabúdať na všetky biologické, chemické, fyzikálne i matematické aspekty tejto úlohy. Pri hlbšom skúmaní tejto oblasti narazíme aj na mnohé problémy. Hlavným negatívom ostáva zložitosť polypeptidových štruktúr, čoho následkom nemôžeme presne zachytiť všetky fyzikálne aspekty pôsobiace na energetickú funkciu. Pracujeme len zo zjednodušenými polypeptidovými modelmi a z tohto dôvodu nemôžeme zaručiť zhodu optimálnej štruktúry z reálnym stavom proteínu.

Autorkinou snahou je zrozumiteľne priblížiť túto zložitú, stále aktuálnu, problematiku čitateľovi a poukázať na metódy globálnej optimalizácie, ktoré boli nejakým spôsobom aplikované na úlohu skladania proteínov.

Úvodná časť je venovaná globálnej optimalizácii, jej matematickému zápisu, najznámejším algoritmom a v neposlednom rade jej využitiu pri riešení praktických úloh. Jadro práce tvorí už spomínaný problém skladania proteínov, kde sú v úvode stručne vysvetlené základné pojmy a používané metódy. Nasleduje priblíženie modelov, ktorými popisujeme proteínovú štruktúru. Základ tvoria algoritmy globálnej optimalizácie používané pri riešení a popis ich fungovania. V závere sú stručne zhrnuté a porovnávané použité metódy.

Kapitola 2

Globálna optimalizácia

2.1 Extrémy funkcií

Optimalizačné úlohy, pri ktorých hľadáme najlepšiu možnosť spomedzi niekoľkých variant, vedú mnohokrát k výpočtu extrémov funkcie. Pri reálnych funkciách rozlišujeme dva druhy extrémov: minimum a maximum, pričom každý z nich má lokálny alebo globálny charakter.

V úvode tejto kapitoly si stručne pripomenieme základné definície, týkajúce sa extrémov funkcií.

Definícia 1 *Funkcia $f : M \to \mathbb{R}$, kde $\emptyset \notin M \subset \mathbb{R}^n$ má v bode $\boldsymbol{x}^* \in M$* ***lokálne maximum****, ak existuje okolie bodu $\boldsymbol{x}^*$ také, že pre všetky $\boldsymbol{x}$ z tohto okolia platí $f(\boldsymbol{x}) \leq f(\boldsymbol{x}^*)$.*

Definícia 2 *Funkcia $f : M \to \mathbb{R}$, kde $\emptyset \notin M \subset \mathbb{R}^n$ má v bode $\boldsymbol{x}^* \in M$* ***ostré lokálne maximum****, ak existuje okolie bodu $\boldsymbol{x}^*$ také, že pre všetky $\boldsymbol{x}$, s výnimkou bodu $\boldsymbol{x} = \boldsymbol{x}^*$, z tohto okolia platí $f(\boldsymbol{x}) < f(\boldsymbol{x}^*)$.*

Lokálne minimum sa zavádza podobne.

Definícia 3 *Funkcia $f : M \to \mathbb{R}$, kde $\emptyset \notin M \subset \mathbb{R}^n$ má v bode $\boldsymbol{x}^* \in M$* ***globálne maximum****, ak pre všetky $\boldsymbol{x} \in M$, platí $f(\boldsymbol{x}) \leq f(\boldsymbol{x}^*)$. Funkcia f má v bode $\boldsymbol{x}^*$* ***ostré globálne maximum****, ak pre všetky $\boldsymbol{x} \in M$, s výnimkou bodu $\boldsymbol{x} = \boldsymbol{x}^*$, platí $f(\boldsymbol{x}) < f(\boldsymbol{x}^*)$.*

Globálne minimum sa zavádza podobne.

Definícia 4 *Ak má funkcia $f : M \to \mathbb{R}$, kde $\emptyset \notin M \subset \mathbb{R}^n$ v bode $\boldsymbol{x}^* \in M$ maximum, nazývame $\boldsymbol{x}^*$* ***bodom maxima funkcie*** *f. Funkčnú hodnotu $f(\boldsymbol{x}^*)$ nazývame potom* ***maximálnou hodnotou funkcie****.*

2.2 Úloha globálnej optimalizácie

Optimalizáciou rozumieme problém nájdenia globálneho extrému *účelovej funkcie* $f(x)$ definovanej na *stavovom priestore* a bodu, v ktorom tohto extrému dosiahne. Definičný obor funkcie f je daný sústavou obmedzení, ktoré sú väčšinou realizovateľné sústavou rovníc. Body z definičného oboru nazývame *prípustne riešenia.*

Základnú úlohu globálnej optimalizácie môžeme formálne vyjadriť ako

$$\min f(x) \tag{2.1}$$

$$\text{za podmienky } x \in D$$

$$D = \{x : l \leq x \leq u; g_k(x) \leq 0; k = 1, \ldots, J\}$$

Kde $x \in R^n$ je reálny $n-$rozmerný vektor *rozhodovacích premenných*
$f : R^n \to R$ je *účelová funkcia*
$D \subset R^n$ je neprázdna *množina prípustných riešení*
l, u sú horné a dolné hranice pre x
$g : R^n \to R^n$ je vektorová funkcia.

Pokiaľ je f diferencovateľná, môžeme k hľadaniu extrému použiť diferenciálny počet a jednoducho z nájdených lokálnych extrémov vybrať ten najväčší.

Poznámka 1 *Pokiaľ hľadáme maximum funkcie $f(x)$ a máme k dispozícii len algoritmus pre hľadanie minima, nájdeme minimum funkcie $-f(x)$, ktoré funkcia dosiahne v tom istom bode ako maximum $f(x)$. V ďalšom texte, ak sa budeme zaoberať hľadaním minima, vieme nájsť aj maximum a opačne.*

Ak však máme zadanú všeobecnú funkciu definovanú na všeobecnej množine, situácia je omnoho zložitejšia. Funkcia nemusí byť spojitá, množina prípustných riešení môže byť zložito zadaná, nespojitá. Aj keď vieme nájsť lokálne extrémy, ich počet môže byť veľmi vysoký. Účelová funkcia je často definovaná v mnohorozmernom priestore a nie sme schopní riešenie v rozumnom čase nájsť.

O zložitosti úlohy teda rozhoduje mnoho faktorov, medzi hlavné patrí spojitosť účelovej funkcie, počet jej lokálnych extrémov na množine prípustných riešení a ich rozmiestnenie, a ohraničenosť množiny prípustných riešení.

2.3 Metódy globálnej optimalizácie

V literatúre sa môžeme stretnúť s rôznymi spôsobmi delenia metód použiteľných k riešeniu problému globálnej optimalizácie. Napríklad rozdelenie **podľa historického vývoja a jednotlivých disciplín**: štruktúrna analýza (1930), teória hier (1944), lineárne programovanie (1947), operačný výskum (1950), teória hromadnej obsluhy (1951), nelineárne programovanie (1951), optimalizácia zásob (1951), dynamické programovanie (1957), sieťová analýza (1958), celočíselné programovanie (1958), viackriteriálna optimalizácia (1970).

Ďalším spôsobom je **rozdelenie podľa toho či ide o statické alebo dynamické modely**. Väčšina modelovaných systémov sa vyvíja v čase. Ak je tento vývoj pre skúmaný účel zanedbateľný, použijeme model v ktorom čas nevystupuje, hovoríme o statickom modeli. Model, v ktorom čas explicitne vystupuje, nazývame dynamický. Pokiaľ čas považujeme za premennú, ktorá nadobúda hodnôt z určitého intervalu, vravíme, že ide o model so spojitým časom. Pokiaľ nás však zaujímajú len určité časové okamžiky, tzn. čas ako premenná nadobúda len konečný počet hodnôt, ide o model s diskrétnym časom.

Inou možnosťou je **rozdelenie podľa toho či ide o deterministické alebo stochastické modely**. Deterministický je model, ktorý nepoužíva pravdepodobnosť a teda ani nepostihuje prípadne náhodné vplyvy na systém. O stochastických modeloch rozprávame hlavne vtedy, keď model bez použitia pravdepodobnostného aparátu nieje adekvátnym zobrazením reality. Typickými stochastickými modelmi sú modely z oblasti teórie hromadnej obsluhy, či niektoré modely optimálneho riadenia zásob.

Ďalšie spôsoby rozdelenia metód globálnej optimalizácie môžeme nájsť napr. v [25].

Stratégie a možnosti riešenia úloh globálnej optimalizácie môžeme ďalej rozdeliť do dvoch skupín:

1. exaktné, presné metódy
2. heuristické metódy

Algoritmy patriace do prvej skupiny na základe teoretických základov vedú k nájdeniu globálneho extrému. Do tejto skupiny zaraďujeme deterministické modely. Ich nevýhodou zostáva, že pri viacdimenzionálnych

problémoch či komplikovanejšej účelovej funkcii sa výpočet stáva neprimerane náročným. Práve v týchto prípadoch sa odporúča siahnuť po metódach heuristických, ktoré však teoreticky nezaručujú konvergenciu. Heuristické algoritmy sú väčšinou založené na analógiách z prírody a rôznych vedných disciplín a využívajú stochastické metódy. Použitím vhodného algoritmu na vhodné dáta sa môžeme s vysokou pravdepodobnosťou dopracovať k dobrému výsledku. Preto je vždy potrebné zvážiť čo od výpočtu očakávame. Pokiaľ nám vyhovuje aproximovaný výsledok s akceptovateľnou pravdepodobnosťou za kratší čas, sú heuristické metódy vhodné, netreba však zabúdať na ich nevýhody.

Práve na tomto rozdelení metód si spomenieme základné algoritmy a postupy. Cieľom je čitateľovi poskytnúť náhľad do týchto metód, prehľad najpoužívanejších postupov, nie podrobný popis a vysvetlenie. Pre detailnejší prístup k danej metóde odporúčam napr. *Journal of Global Optimization*, či inú dostupnú literatúru, ktorej je v súčasnej dobe veľké množstvo.

2.3.1 Exaktné metódy

Medzi exaktné metódy patria:

1. ***Prirodzené prístupy (naive approaches)***, ktoré zahŕňajú dobre známe pasívne paralelné alebo priame postupné metódy globálnej optimalizácie ako jednoduchá mriežka (uniform grid), pokrývanie priestoru (space covering) a jednoduché náhodné prehľadávanie (pure random search). Tieto metódy konvergujú aj za slabých predpokladov, ale zpravidla sa nepoužívajú na viacdimenzionálne náročnejšie problémy.

2. ***Kompletné enumeratívne prehľadávacie stratégie (enumerative search strategies)*** sú založené na kompletnom usmerňovanom prechádzaní všetkými možnými riešeniami. Sú vhodné pri riešení kombinatorických problémov, či úloh konkávneho programovania.

3. ***Dráhové metódy a homotópne metódy (homotopy and trajectory methods)*** sú založené na prechádzaní všetkými stacionárnymi bodmi účelovej funkcie, čo vedie k zoznamu všetkých (lokálnych aj globálnych) extrémov. Patria sem algorimty založené na diferenciálnych rovniciach (differential equation model), prehľadávacie stratégie sledovania cesty (path-following search strategy), a tiež známe metódy pevného bodu (fixed-point methods) a pivot algoritmus (pivoting algoritmus).

4. ***Relaxačné metódy (relaxation methods)*** nahrádzajú počiatočný optimalizačný problém sekvenciou podproblémov, ktoré sú výpočetne menej náročné. Medzi metódy používané k zjemňovaniu problému tak aby aproximoval počiatočné zadanie patria rezné roviny(cutting planes), konštrukcia diverznej minorantnej funkcie (diverse minorant function construction), dekompozičné stratégie (dekompozition strategy) a pod. Tieto metódy sa využívajú pri diverznej štruktúre úlohy globálnej optimalizácie.

5. ***Algoritmus vetví a medzí (Branch and bound algorithm)***, kde v prvej fáze vetvenia množinu prípustných riešení nahradíme niekoľkými menšími množinami, a túto procedúru opakujeme rekurzívne. V druhej fáze hľadáme dolné a horné hranice pre optimálne riešenie v rámci množiny prípustných riešení. Použitie tohto algoritmu je podmienené bližšou znalosťou štruktúry danej úlohy (napr. Lipschitzovskú konštantu pre funkcie f a g, hladkosť funkcií a pod.). Používa sa na značné množstvo optimalizačných úloh, napr. kombinatorická optimalizácia, konkávna minimalizácia, Lipschitzovské optimizačné úlohy apod..

6. ***Bayesov prehľadávací algoritmus (Bayesian search algorithmus)*** je založený na apriórnom predpoklade stochastického modelu, štatistickej informácie o skupine funkcií, z ktorej f pochádza. Hlavným problémom sa ukazuje hlavne výber vhodného štatistického modelu, pričom sa musí prihliadať na konkrétnu aplikáciu problému. Tento algoritmus nieje vhodný pri riešení viacrozmerných úloh, aj keď v súčasnosti už existujú vylepšenia, ktoré umožňujú využiť tento princíp aj v týchto úlohách.

7. ***Adaptívny stochastický prehľadávací algoritmus (Adaptive stochastic search algorithmus)*** funguje na náhodnom výbere z množiny D. Na vylepšenie tejto metódy sa používajú napr. jednoduché zoskupenie (sample clustering), štatistické stop pravidlo (statistical stop rule) a pod. Jeho výhoda je v použiteľnosti, ako na úlohy diskrétne tak spojité, pri veľmi obecných predpokladoch.

8. ***Integrálne metódy (Integral methods)*** sú založené na určení esenciálneho supréma úžitkovej funkcie f na množine D, aproximovaním množín s rovnakou funkčnou hodnotou f.

2.3.2 Heuristické metódy

K týmto metódam patria:

1. ***Rozšírenie metódy lokálneho prehľadávania (Extensions of local search methods)*** je čiastočne heuristická metóda často využívaná v praxi. Základná myšlienka spočíva v počiatočnom použití mriežkového prehľadávania (grid search), alebo použití iného globálneho prehľadávania. Následne sú použité vhodné lokálne metódy. Opäť je k dispozícii niekoľko vylepšení základného algoritmu.

2. ***Evolučné metódy, genetický algoritmus (Evolution strategies, Genetic algorithms)*** sú metódy, ktoré, ako je už z názvu jasné, sa inšpirovali biologickou evolúciou, procesom prirodzeného výberu a prežitím najlepšieho jedinca. Kombinuje adaptívny a evolučný vývoj s funkčnou optimalizáciou. Evolučné techniky patria medzi najrýchlejšie sa rozvíjajúce obory, mnohokrát sa vyčleňujú ako samostatná skupina algoritmov. Výhodou je rýchlosť prehľadávania stavového priestoru a to, že genetické algoritmy nepotrebujú pre svoju činnosť žiadne dodatočné informácie, stačí len funkčná hodnota v ľubovoľnom bode. Okrem genetických algoritmov do skupiny evolučných metód patria napr. metóda mravčích kolónií (ant colony optimization) alebo metóda imunologického systému (immunology system method).

3. ***Metóda simulovaného žíhania (Simulated Annealing)*** vychádza z fyzikálnej predstavy procesov, ktoré prebiehajú pri odstraňovaní defektov kryštalickej mriežky. Kryštál sa zahreje na vysokú teplotu a potom sa pomaly ochladzuje (žíha). Patrí do skupiny algoritmov, ktoré využívajú sekvenčné náhodné prehľadávanie. Pri žíhaní sa sústava snaží dostať do optimálneho stavu, ktorý zodpovedá mininálnej energii, a teda aj kryštálu bez defektu. Simulované žíhanie je jedným z najúspešnejších tradičných stochastických optimalizačných algoritmov, ktorý je mnohokrát rýchlejší a aj presnejší ako genetické algoritmy.

4. ***Metódy zakázaného prehľadávania (Tabu search)*** vychádzajú z myšlienky zakázania prehľadávania bodov, ktoré sme už navštívili pred niekoľkými krokmi. Počas výpočtu sa udržuje tzv. zakázaný zoznam, ktorý zabraňuje zacykleniu a tým zvýši globálnosť algoritmu. Tento zoznam sa používa aj ku konštrukcii modifikovaného suseda. Tieto metódy sa používajú hlavne na riešenie kombinatorických optimalizačných úloh (napr. rozvrhovací problém, problém obchodného cestujúceho a pod.).

5. ***Metóda aproximácie konvexného podhodnoteného odhadu (Approximate convex underestimation)*** je založená na snahe odhadnúť hlavné konvexné charakteristiky úžitkovej funkcie. Pri správnej voľbe dáva algoritmus veľmi dobré výsledky, používa sa hlavne na hladké problémy globálnej optimalizácie.

6. ***Naväzovacia metóda (Continuation method)*** najprv transformuje úžitkovú funkciu na hladšiu, jednoduchšiu funkciu, ktorá má menej lokálnych extrémov, potom sa použije procedúra lokálnej optimalizácie k načrtnutiu minima pôvodnej funkcie.

7. ***Sekvenčné zlepšenie lokálneho optima (Sequential improvement of local optima)*** je založené na konštrukcii pomocnej funkcii k postupnému hľadaniu lepšieho optima. Patria sem metódy tunelovanie (tunneling), vypustenie (deflation) a pod..

Kapitola 3

Globálna optimalizácia v praxi

Úlohy z rôznych odvetví, v ktorých ide o výber najlepšej alternatívy z danej množiny variant, mnohokrát vedú k výpočtu extrémov funkcií viacerých premenných. Hľadanú najlepšiu možnosť potom reprezentuje minimum alebo maximum tejto funkcie. Motivácie zo začiatku rozvoja optimalizácie prichádzali hlavne z oblasti fyziky, neskôr z rôznych oblastí, napr. technických, ekonomických, vojenských či lekárskych. Ale až rozvoj počítačov v päťdesiatych rokoch zaviedol pozornosť na realizáciu výpočtov úloh, ktorým predtým bránil veľký počet premenných či obmedzujúcich podmienok. Od tej doby sa hlavná pozornosť v aplikovanej optimalizácii venuje vývoju nových účinnejších algoritmov či heuristík.

V súčasnosti sa s riešením optimalizačných úloh stretávame v rôznych odvetviach. Masívny rozvoj optimalizácie zabezpečili najmä úlohy z oblasti ekonómie, v ktorej sa extremalizačné problémy vyskytujú v mnohých praktických problémoch. Jednu oblasť tvoria problémy *optimalizácie výrobných programov*. Tu typicky nezáleží, z hľadiska tvorby správneho modelu, či ide o menšiu firmu , či celé odvetvie, ani o aký podnik ide. V týchto úlohách podnik hľadá správne kvantifikované rozhodnutie napr. objem investícii pre určité obdobie, vyhovujúci výrobný program. Pekný príklad k tejto tématike môžeme nájsť napríklad v [3] str.126, kde je publikovaný známy problém konvexnej optimalizácie : Optimálne využitie mlieka v Holandsku(1960). Z mlieka sa vyrábajú 4 produkty s ročnou produkciou: x_1 - mlieko, x_2 - maslo, x_3 - tvaroh, x_4 - syr, a cenami na trhu p_1 až p_4. Úlohou je optimálne rozvrhnúť výrobu za rôznych podmienok, pričom nám ide o maximalizáciu zisku.

Ďalšou úlohou z ekonomického prostredia tvoria *dopravné problémy*. Tieto problémy sa odvádzajú od základného dopravného problému, v ktorom

sa rieši splnenie požiadaviek geograficky rozptýlených zákazníkov, ktorých požiadavky sú zabezpečované z jedného skladu za určitých obmedzení. Zaujíma nás taký plán prepravy, ktorý minimalizuje buď čas dopravy, alebo jej náklady. Medzi úlohy, ktorých základ tvorí práve základný dopravný problém patrí napr. vyzdvihnutie pošty zo schránok alebo mincí z automatov, obchádzka lekára po pacientoch, cestovanie predajcu a pod..

V *alokačných problémoch* hľadáme optimálne umiestnenie výrobných stredísk, skladov a pod. aby sa minimalizovali náklady na dopravu či kooperáciu.

Distribučné problémy sa zas zaoberajú rozvrhnutím výroby na rôzne typy strojov za účelom zefektívnenia výroby.

Ďalšiu skupinu exremalizačných úloh tvoria úlohy z technických oblastí. Veľkú skupinu tvoria optimalizačné *úlohy z oblasti fyziky.* V [3] str.153 nájdeme zadania niektorých úloh : napr. princíp najmenšieho času pre svetelné lúče, ktorý vraví o tom, že svetlo z bodu a do bodu b cestuje trasou, ktorá minimalizuje čas. Optimalizácia sa využíva aj pri princípe rovnovážnej pozície mechanických systémov a pod..

Množstvo úloh, ktoré je možné previesť na úlohy globálnej optimalizácie nájdeme aj v chémii. Jedným z najznámejších, spomenutých aj v knihe [14], je výpočet chemickej rovnováhy v plynnej sústave. Zložitejšie sú už úlohy z chemického inžinierstva, napr. pri analýze Gibsovej voľnej energie (maximálne množstvo energie, ktoré môžeme získať zo systému bez zmeny objemu a pridania tepla), ktorá dosahuje minima pri rovnovážnom stave.

Z biologicko-chemického hľadiska sú zaujímavé príklady zo simulácii molekulovej dynamiky, kde sa v úvode každej simulácie hľadá minimum energie simulovaného systému, či problém skladania proteínov, ktorý tiež vychádza z minimalizácie energie.

Zaujímavý príklad využitia niektorých metód globálnej optimalizácie nájdeme aj v článku [26], kde autori problém lokalizácie aktívnych centier mozgu previedli na problém hľadania minima úžitkovej funkcie.

3.1 Konkrétne príklady využitia globálnej optimalizácie v praxi

V tejto časti sa pozrieme na niektoré zaujímavé príklady využitia globálnej optimalizácie pri riešení konkrétnych už publikovaných problémov.

3.1.1 Využitie metód globálnej optimalizácie v medicíne

Techniky aplikovanej globálnej optimalizácie sa ukázali efektívne v mnohých lekárskych otázkach ako napr. diagnostika, plánovanie liečby, s ktorým súvisí identifikácia rizík a ich zahrnutie do plánu, moderné vyšetrovacie techniky, a pod. Niektoré z týchto problémov boli riešené aj v [16]. Autori sa špeciálne venovali otázke využitia metód optimalizácie pri diagnostike rakoviny prsníka, dynamike ľudského mozgu a predpovede epileptických záchvatov, plánovanie liečebných procedúr na príklade rádioterapie, či pri zobrazovacích problémoch.

Problém diagnostiky

K dispozícii máme súbor dát od rôznych pacientov, ktoré zahŕňajú rôzne parametre (vek, teplota, krvný tlak, veľkosť nádoru a pod.), pričom poznáme z toho vyplývajúcu diagnózu. Prirodzený spôsob diagnostiky nového pacienta matematickými metódami je teda použiť tieto známe dáta so známou diagnózou (tzv. trénovacia množina) pri konštrukcii modelu, ktorého úlohou bude klasifikovať príčiny ochorenia na základe známych informácií. Matematicky máme množinu o N prvkoch, a každý z týchto prvkov má konečný počet znakov, ich počet je n. Teda každý prvok množiny môžeme zapísať ako dvojicu (x_i, y_j), $i = 1, \ldots, N$, kde $x_i \in R^n$ je $n-$rozmerný vektor a y_i je trieda znakov. Teda y_i určuje, do ktorej skupiny prvok patrí a je apriórne známe v trénovacej množine. Na základe trénovacej množiny chceme nový prvok správne zaradiť a teda priradiť mu triedu y. Snažíme sa vlastne vyriešiť problém hľadania optimálnych hodnôt parametrov v modeli. V tomto prípade sa používa jednoduchý geometricky prístup, ktorý sa neskôr prispôsobuje zložitosti zadaného problému. Základ tohto prístupu je reprezentácia každého prvku množiny ako bodu v $n-$rozmernom priestore. Tieto body potom môžeme geometricky oddeliť plochou. Takže problém diagnostiky môžeme pretransformovať na problém nájdenia geometrických parametrov oddeľovacích plôch. Tieto parametre nájdeme ako riešenie optimalizačného problému minimalizácie chýb nesprávneho zaradenia prvkov trénovacej množiny. Po identifikovaní týchto parametrov, každý nový prvok môžeme podľa jeho geometrickej lokácie zaradiť do skupiny. Ako príklad tohto problému je v [16] uvedená diagnostika rakoviny prsníka, kde trénovacia množina obsahuje 569 pacientov a na každom pozorovaných 30 znakov, pričom je dané či nádor bol rakovinový alebo nie. Práve na tomto príklade sú ilustrované vhodné postupy a ich vylepšenia.

Predpovede rizika

V tejto oblasti ide o odhadnutie možného rizika, ktoré hrozí pacientovi, na základe lekárskych pozorovaní. Jednou z možností riešenia je využitie tzv. logickej analýzy dat (Logical Analysis of Data), v ktorej ide o kombináciu optimalizácie a Booleovej logiky. Matematicky $\Omega \subset R^n$ je množina pozorovaní, pričom Ω^+ sú pozitívne pozorovania a Ω^- sú pozorovania negatívne. Tiež definujeme akýsi vzor správania $P = \{x \in R^n; x_i \leq \alpha_i (i \in I), x_j \geq \beta_j (j \in J); I, J \subseteq \{1, \ldots, n\}\}$, kde $\alpha_i, i \in I$, $\beta_j, j \in J$, sú reálne čísla. Vzor P je pozitívny ak $P \cap \Omega_+ \neq \emptyset$ a $P \cap \Omega_- = \emptyset$. Vzor je charkterizovaný parametrami stupeň, rozsah, a riziko. Stupeň je daný počtom nerovností, ktoré definujú tento vzor, rozsah je počet pozorovaní vo vzore P, riziko vzoru P je dané ako $\rho_P = \frac{|P \cap \Omega_+|}{|P \cup \Omega|}$. Na základe týchto vlastností sme schopní rozdeliť ich na nízkorizikové a vysokorizikové vzory. Tento prístup bol ilustrovaný na príklade predpovede rizika úmrtnosti na koronárne problémy (infarkt a pod.) a zaradenia pacientov do rizikových skupín.

Mozgová dynamika a predpoveď epileptických záchvatov

Ľudský mozog je jeden z najzložitejších komplexov, ktoré boli kedy skúmané. Enormný počet neurónov a ich dynamické prepojenie pridáva na zložitosti problémov. Pokrok v tejto oblasti je spojený najmä s používaním elektroencefalogramu (EEG), ktorý nám poskytuje informácie o funkciách mozgu. Spojením medicíny, teórie nelineárnej dynamiky a matematiky sa vedci pokúšajú porozumieť jeho fungovaniu. V úvode tejto tematiky sa v [16] venujú priblíženiu mozgovej dynamiky pomocou teórie dynamických systémov a zavedeniu odhadu maxima Lyapunovho exponentu-SPL (, ktorý je definovaný ako miera divergencie fázových trajektórií , čiže v prípade obecnej trajektórie určuje, ako veľmi sa v priebehu doby trajektórie vychádzajúce z veľmi blízkych počiatočných bodov od seba líšia). Získame teda odhad SPL pre jednotlivé body v mozgovej kôre, a s blížiacim sa záchvatom môžeme pozorovať zmeny. Dôležité je si uvedomiť, že pokiaľ v jednotlivých miestach prebieha podobná operácia, potom hodnota SPL v týchto bodoch konverguje k rovnakej hodnote. Tieto body potom nazývame kritickými. Zhruba hodinu pre záchvatom môžeme pozorovať pomocou elektród na kritických miestach charakteristické zmeny. Práve výber vhodných kritických miest má vplyv na kvalitný odhad blížiaceho sa záchvatu. Na tento výber sa používajú práve metódy kvadratickej optimalizácie.

Plánovanie liečby

Najväčším oddelením kde sa využívajú optimalizačné techniky pri plánovaní liečby je ožarovacia terapia. Táto metóda, na liečenie rakoviny pomocou vysokoenergetickej radiácie, zamedzuje rakovinovým bunkám sa ďalej šíriť. Rozlišujeme dva typy ožarovania : externé lúčové ožarovanie, alebo tzv. brachyterapia, kde sú rádioaktívne zdroje umiestnené priamo do nádoru alebo jeho blízkosti. Prvým krokom každej plánovacej metódy je rozbor snímok pacienta na jednotlivé pixely: kritické (zdravé tkanivo citlivé na ožarovanie), telo (zdravé tkanivo nie veľmi citlivé na ožarovanie) a nádor (rakovinové bunky). Plánovanie externej rádioterapie zahŕňa nastavenie smeru , intenzity a tvaru lúču, pričom potrebujeme optimalizovať dávku pre rakovinovú oblasť a zároveň minimalizovať poškodenie zdravých orgánov. V tomto prípade boli použité rôzne techniky: lineárne, nelineárne programovanie, a pod. V [16] sú stručne popisované niektoré už publikované metódy pre danú problematiku.

3.1.2 Optimalizačné problémy akustiky v oceánoch (podmorskej akustiky)

Tento zaujímavý problém bol riešený v [24]. Podmorská akustika sa zvyčajne používa k lokalizácii rôznych objektov ako ponorky či húfy rýb. Novšími sú už aplikácie k preskúmaniu oceánu samotného, jeho teploty, zloženia jeho podložia, vlastnosti morského dna, či výskytu ľadovcov.

Jednou z dôležitých otázok tejto problematiky je správny prístup ku každému problému tak, aby bola zaručená jedinečnosť riešenia. Z matematického hľadiska je hlavým problémom rozľahlý priestor neznámych a jeho nekonvexnosť.

Jendou z aplikácií je práve rekonštrukciou 3-D teplotného profilu na základe šírenia zvukových vĺn a ich fyzikálneho správania sa vo vodnom prostredí. Na základe meraní akustickej energie sa snaží odhadnúť environmentálne parametre skúmaného prostredia (napr. hrubosť povrchu dna, odrazivosť, teplotu, hustotu, pórozitu, ...).

Odhad environmentálnych parametrov zahŕňa štúdium efektov veľkého množstva možných hodnôt týchto parametrov. K dispozícii máme experimentálne dáta a model, ktorého výstupy pri znalosti správnych hodnôt environmentálnych sú rovnaké ako namerané dáta. Základným problémom teda ostáva, ako zo známych experimentálnych dát a znalosti modelu, určiť správne hodnoty jeho parametrov. Teda rieši sa inverzný problém pričom sa používa buď funkcia, ktorá vyjadruje zhodu medzi modelom a dátami (táto funkcia sa optimalizuje), alebo prehľadávanie celého priestoru hodnôt parametrov.

Autor v [24] popisuje aj prehľadovo niekoľko metód, ich výhody aj nevýhody, či návrhy na ich vylepšenie v špecifických príkladoch. Zoberá sa simulovaným žíhaním, genetickým algoritmom, dizajnérskym algoritmom a jeho linearizáciou. Podrobne sa zaoberá novým spôsobom porovnávania dát s modelom MFP (Matched field processing).

V ďalšej časti sa budem bližšie venovať konkrétnému, stále aktuálnemu problému, skladania proteínov, ktorý je ekvivalentný problému hľadania minima funkcie potenciálnej energie daného polypeptidového systému.

Kapitola 4

Problém skladania proteínov (Protein folding problem)

Problém skladania proteínov a teda minimalizovania funkcie potenciálnej energie môžeme matematicky sformulovať ako minimalizáciu nelineárnej funkcie $f(x)$, ktorá má viacnásobné lokálne minimá, pre x patriace do D, definovanej dolnými a hornými hranicami pre každý parameter x_i.

Pre pochopenie a lepšiu orientáciu v použitých metódach globálnej optimalizácie je dôležité pripomenúť základné a používané pojmy a termíny z oblasti proteínov, ich skladania, a fyzikálneho pozadia tejto problematiky. V tejto časti sa tiež podrobnejšie oboznámime s touto preblemtikou.

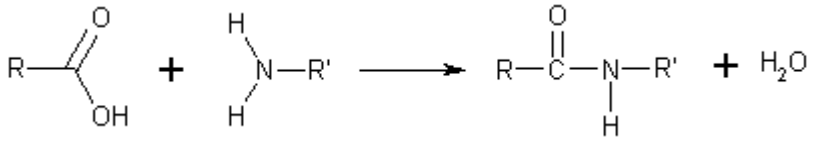
4.1 Základné pojmy

4.1.1 Proteíny

Proteíny sú polyméry vytvorené z reťazca monomérnych jednotiek - aminokyselín, vzájomne pospájaných peptidovou väzbou CO-NH (väzba vzniká medzi molekulami, reakciou medzi aminoskupinami -NH^2 jednej molekuly a karboxylovými skupinami -COOH druhej, s uvoľnením vody H_2O).

Po pridaní do polypeptidového reťazca aminokyselinu nazývame *jadro (residue)*, príležitostne sa značí R a spojenú skupinu atómov uhlíka, dusíka a kyslíka nazývame *hlavný reťazec*.

Poradie aminokyselín v polypeptidovom reťazci (bielkoviny, proteíny) je určené geneticky v DNA, kde trojica nukleotidov kóduje jednu aminokyselinu.



Obrázok 4.1: Príklad reakcie pri ktorej vzniká peptidová väzba

Štruktúra proteínov

Väčšina proteínov má 3 - dimenzionálnu štruktúru. Tvar, do ktorého sa proteín prirodzene zloží nazývame *prirodzený, počiatočný stav*. Niektoré proteíny sa skladajú samostatne na základe chemickej štruktúry, mnohé sa skladajú za pomoci molekulárnych *chaperonov*.

Z hľadiska štruktúry sa rozlišuje niekoľko úrovní pohľadu na proteíny:

- ***Primárna štruktúra**, sekvencia* je určená poradím aminokyselín v polypeptidovom reťazci.

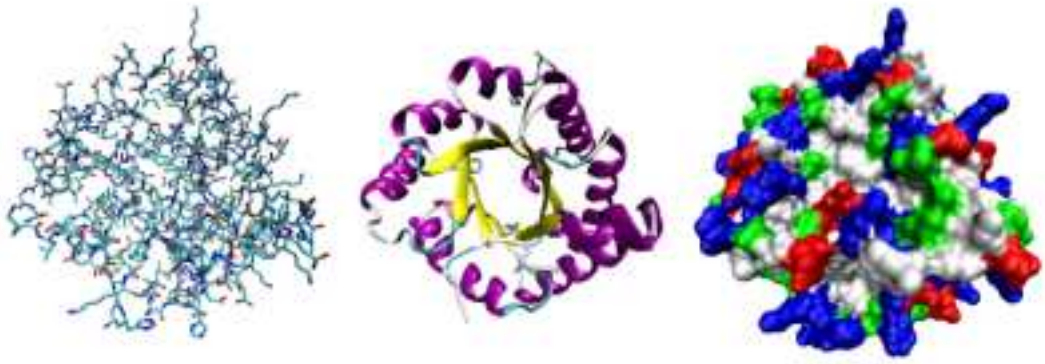
- ***Sekundárna štruktúra*** je označenie pre priestorové usporiadanie peptidového reťazca, ktoré vzniká vďaka vodíkovo - mostíkovým väzbám medzi atómmi polárnych skupín C=0 a H-N. Najznámejšie sekundárne štruktúry sú *alfa-hélix* (pravotočivá závitnica) a *beta-štruktúra* (skladaný list beta). Sekundárna štruktúra je lokálna, takže v rámci jednej proteínovej molekuly sa môže vyskytnúť viacero sekundárnych štruktúr.

- ***Terciárna štruktúra*** je označenie pre priestorové usporiadanie peptidového reťazca, ktoré vzniká vodíkovo - mostíkovými väzbami, iónovou väzbou, disulfidovými väzbami či nepolárnymi Van der Waalsovými silami medzi jednotlivými sekundárnymi štruktúrami. Jej výsledkom je buď *fibrilárna podoba* (vzniká väzbami medzi rôznymi polypeptidovými reťazcami) alebo *globulárna podoba* (vzniká väzbami medzi časťami toho istého reťazca) bielkoviny.

- ***Kvartérna štruktúra*** je označenie pre štruktúru stabilného *oligoméru* (čiže istej skupiny viacerých peptidových reťazcov).

4.1.2 Skladanie proteínov

Ako sme už spomínali skladanie je fyzikálny proces, pri ktorom sa molekuly proteínov skladajú do svojej charakteristickej 3 - dimenzionálnej štruktúry,

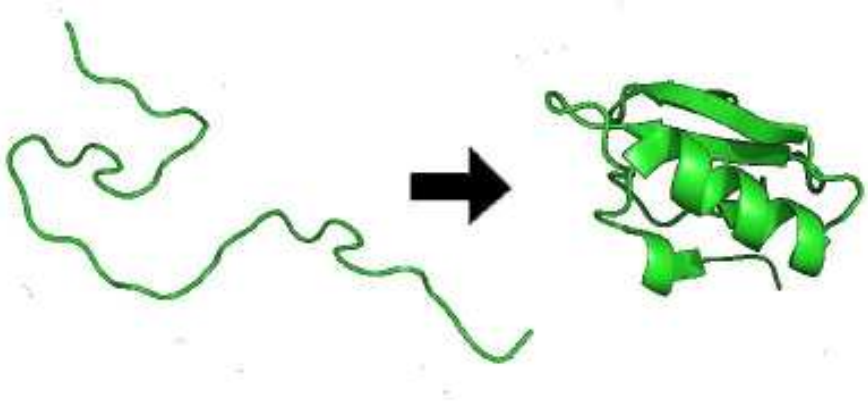


Obrázok 4.2: Tri pohľady na štruktúru monoméru proteínu tri-fosfát izomerázy. Ľavý pohľad na všetky atómy, farebne rozlíšené. Uprostred sú farebne rozlíšené podľa sekundárnej štruktúry, napravo farebne rozlíšené podľa typu jadra (acidové jadrá červená, bazické jadrá modrá, polárne jadrá zelené, nepolárne jadrá biele).[1]

prirodzeného stavu. Proteíny sa syntetizujú ako lineárny reťazec zložený až z niekoľkých stoviek aminokyselín. Práve poradie aminokyselín (primárna štruktúra) podmieňuje do akej jedinečnej štruktúry sa proteín zloží. Skladanie proteínu trvá v rozsahu niekoľkých milisekúnd až sekundy, pričom štruktúra prechádza aj niekoľkými medzištádiami. Väčšina zložených proteínov obsahuje tzv. *hydrofobické jadrá*. Hydrofobický efekt znamená, že nepolárne molekuly sa držia pri sebe vo vodnom roztoku, extrémny prípad je olej vo vode. Hydrofobické aminokyseliny teda tvoria tieto jadrá, na povrchu ktorých sa nachádzajú bočné reťazce, ktoré sú vystavené pôsobeniu vodného roztoku, v ktorom sa proteíny v organizmoch vyskytujú. Všeobecne sa dá povedať, že proteíny sa skladajú tak, aby sa minimalizoval počet bočných reťazcov, a tým sa minimalizovala interakcia s molekulami vody.

Práve chybné skladanie je zodpovedné za napr. Creutzfeldt - Jakobovu chorobu, BSE (chorobu šialených kráv) a Alzheimerovu chorobu. V súčasnosti sa však nevie, prečo dochádza k tomuto chybnému zloženiu, preto sa tejto problematike venuje stála pozornosť.

[1]Zdroj www.wikipedia.com.



Obrázok 4.3: Príklad skladania proteínov.[1]

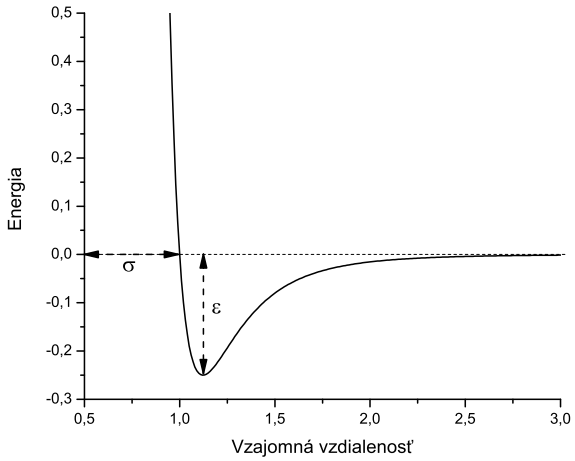
4.1.3 Základné fyzikálné pojmy

Potenciálna energia alebo *polohová energia* (W_p alebo E_p) je druh mechanickej energie, je to práca, ktorú treba vykonať na telese, aby sa z referenčnej energetickej úrovne (tzv.normálneho stavu alebo nulovej hladiny) dostalo do danej novej polohy. Treba si uvedomiť, že potenciálna energia je relatívna, čiže záleží na referenčnej energetickej úrovni, teda na tom, na čo ju vzťahujeme. Pri výpočtoch sa nulová hladina potenciálnej energie volí buď v rovnovážnej polohe, v ktorej sú príslušné sily v rovnováhe, alebo v nekonečne, kde je veľkosť príslušných síl na teleso nulová. Na voľbe nulovej hladiny potenciálnej energie logicky nezáleží - rozhodujúca je iba zmena tejto energie.

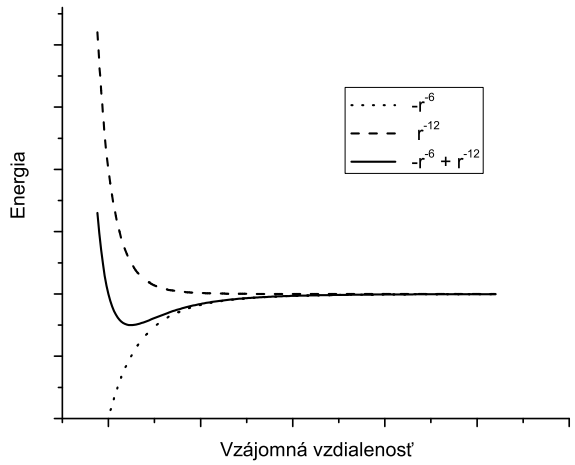
Van der Waalsove sily patria medzi skupinu medzimolekulárnych síl, ktoré vznikajú polarizáciou molekúl na dipóly prípadne multipóly. Zahrňujú príťažlivé aj odpudivé sily. Najznámejšie vyjadrenie Van der Waalsovej potenciálnej energie je *Lennard-Jonesova potenciálna funkcia*, známa tiež ako Lennard-Jonesova 12-6 funkcia, (podrobnosti viď. napr. v [9]). Model, ktorý zobrazuje vzťahy medzi dvoma atómmi vyzerá následovne:

$$V(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{4.1}$$

Kde r je vzájomná vzdialenosť, σ je tzv. *zrážkový priemer* - v tejto vzdialenosti je potenciálna energia nulová, ε je hĺbka potenciálového minima (Viď. Obr.4.4). Výraz r^{-12} popisuje odpudivé vzťahy a r^{-6} popisuje vzťahy príťažlivé (Obr.4.5).



Obrázok 4.4: Lennard-Jonesova potenciálna funkcia



Obrázok 4.5: Priebeh funkcie potenciálnej energie, jej odpudivá a príťažlivá časť

4.2 Predpoveď štruktúry proteínov (Protein structure prediction)

Ako je už z názvu jasné ide o určenie troj-dimenzionálnej štruktúry proteínu na základe znalosti jeho postupnosti aminokyselín, čiže formálne o predpoveď terciárnej štruktúry na základe znalostí primárnej štruktúry.

Existuje však množstvo faktorov, ktoré robia túto úlohu veľmi zložitou, napr. :

- enormné množstvo jednotlivých tvarov, ktoré proteíny môžu nadobúdať
- nedostatok presných znalostí z oblasti fyzikálnej stability danej štruktúry
- mnohé proteíny sa skladajú za prítomnosti pomocných proteínov
- jednotlivé sekvencie sa môžu skladať do rôznych tvarov v závislosti na okolitom prostredí, a tento, síce biologicky aktívny tvar, za normálnych okolností nemusí byť termodynamicky výhodný
- priame simulácie skladania proteínov sú účinné len na malých proteínoch

Avšak vďaka pokroku v počítačovej technike, novým algoritmom dochádza pomaly k prekonávaniu týchto faktorov, a v súčasnosti je realisticky možné predpovedať štruktúry pre menšie proteíny. K tomuto sa používa veľké množstvo rôznych prístupov, ktoré môžeme rozdeliť do dvoch veľkých skupín:

1. ab initio modelovanie (modelovanie z prvotných predpokladov)
2. porovnávacie modelovanie

4.2.1 Ab initio modelovanie

Tieto metódy sa usilujú o modelovanie trojrozmerných bielkovinových modelov od základov, teda model je založený na základných fyzikálnych predpokladoch, a nevychádza z predchádzajúcich znalostí danej štruktúry. Výpočetné procedúry sa buď pokúšajú napodobniť skladanie proteínu, alebo používajú nejakú stochastickú metódu (napr. globálna optimalizácia vhodnej energetickej funkcie). Tieto metódy však vyžadujú obrovské výpočetné zdroje,

a sú teda použiteľné len na krátke bielkovinové reťazce. Pokiaľ daný spôsob chceme využiť pre predpoveď štruktúry rozsiahlejších proteínov, je potreba zefektívniť algoritmy, a rozšíriť výpočetné možnosti, a to buď použitím paralelného počítania alebo reálnejším distribuovaným počítaním. Práve na druhom princípe funguje jeden z najrozšírenejších projektov Folding@home, ktorý distribuovaním programu do domácností simuluje skladanie proteínov či iné dynamické molekulové úlohy.

4.2.2 Porovnávacie modelovanie

Tieto metódy sa opierajú o už vopred vyriešené štruktúry, ktoré považujú za východiska alebo šablóny. Tento postup sa zdá byť veľmi účinný, pretože síce aktuálny počet existujúcich proteínov je obrovský, existuje však len obmedzený počet terciárnych štruktúr do ktorých sa proteín skladá. Predpokladá sa, že v prírode sa vyskytuje asi 2000 zreteľných tvarov proteínového skladania, aj keď existuje mnoho miliónov rôznych proteínov.

Tieto metódy môžeme ešte rozdeliť do dvoch skupín:

- Homologické modelovanie je založené na predpoklade, že dva podobné proteíny budú zdieľať aj podobné štruktúry. Z toho je jasné, že najlepšie výsledky sa dosahujú keď šablóna je čo najviac podobná hľadanému proteínu, čiže ak obsahujú podobné sekvencie aminokyselín. Práve cez podobné postupnosti sa dá určiť aj miera podobnosti proteínov.

- Proteínové reťazenie (Protein threading) porovnáva reťazce aminokyselín neznámej štruktúry s databázou už vyriešených štruktúr. Používa sa aj vyhodnocovacia funkcia, ktorá oceňuje kompatibilitu sekvencie a štruktúry. Táto metóda teda na základe znalostí lineárnej bielkovinového reťazca mu priradí najvhodnejší trojrozmerný model.

Kapitola 5

Používané modely

V tejto kapitole sa bližšie zoznámime z modelmi reprezentujúcimi proteíny v úlohách globálnej optimalizácie, keďže z hľadiska efektivity použitých algoritmov je dôležité vedieť, s ako reálnym modelom daná metóda pracuje.

V súvislosti z modelom je uvedný aj tvar odpovedajúcej funkcie potenciálnej energie.

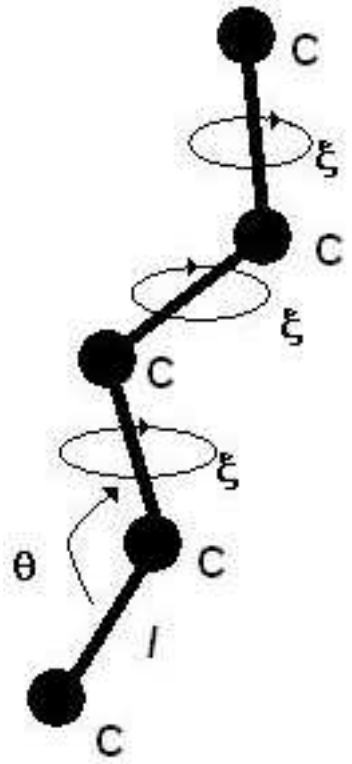
5.1 Jednoduchý hydrofóbno-polárny model (HP model)

Výpočet globálného minima funkcie potenciálnej energie pri reálnom zobrazení proteínu a zahrnutí všetkých energií a ostatných reálnych aspektov nie je ešte v súčastnosti možný, preto pristúpime k vytvoreniu jednoduchého modelu.

Náš model pozostáva z reťazca monomérov C, ktoré sú pospajané väzbami do reťazca. Príslušné vyjadrenie funkcie potenciálnej energie bude zachytávať hlavné sily pôsobiace pri skladaní proteínov. Pozícia každého monoméru (ako môžeme vidieť na obr. 5.2) je vzhľadom k ostatným monomérom v reťazci definovaná premennými parametrami:

1. l - dĺžka väzby
2. θ - väzbový uhol
3. ξ - klinový väzbový uhol

Zafixujeme l, θ, čím znížime počet premenných, ktorými jednoznačne určíme 3-dimenzionálnu štruktúru, na $n-1$. Každá jednotka (monomér) C je *hydrofóbna* (H) alebo *polárna* (P).



Obrázok 5.1: Jednoduchý HP model

Poznámka 2 *V niektorých prípadoch sa stretávame s tým, že pojem klinový väzbový uhol sa používa pre oba typy uhlov vo väzbe.*

Mnohokrát sa pri reprezentácií tohto modelu používa jednoduchá kocková mriežka. Sekvencia je namapovaná na mriežku, každé jadro zaberá jednu bunku na mriežke. Proteín predstavuje teda jednu cestu po mriežke, ktorá sa riadi istými pravidlami. Touto reprezentáciou vystupujú do popredia vzťahy medzi najbližšími jadrami. V literatúre nájdeme aj iný typ reprezentácie, napr. na 2D mriežke [1], diamantovej mriežke [21], či trojuholníkovej mriežke [4].

5.1.1 Reprezentácia kartézskymi súradnicami

Často je potrebné previesť reprezentáciu pomocou vnútorných molekulových súradníc do kartézskych súradnic. Bez ujmy na obecnosti pri tomto pre-

vode volíme pevne tri jednotky. C_1 má súradnice $(0, 0, 0)$, C_2 má súradnice $(-l_2, 0, 0)$ a C_3 má súradnice $(l_3 \cos(\theta_3) - l_2, l_3 \sin(\theta_3), 0)$. Pre C_4 a každú ďalšiu jednotku reťazca hľadáme pozíciu použitím karteziánskej reprezentácie vzhľadom k predchádzajúcim trom jednotkám v reťazci, dĺžke l, uhlom ξ a θ.

Prevod však stále zostáva obšírnym problémom, ku ktorému sa dá pristupovať viacerými spôsobmi, ktoré sú už predmetom chemickej fyziky, viac informácií nájdeme napr. v [23].

5.1.2 Funkcia potenciálnej energie

V súlade z jednoduchosťou tohto modelu, aj funkcia potenciálnej energie E_{total} je charakteristická jednoduchým vyjadrením. Obsahuje 3 komponenty:

1. výraz zahrňujúci energiu medzi rezíduami H-H
2. výraz zachycujúci priestorový odpudivý efekt (odmieta akékoľvek postavenie, ktoré by povoľovalo nerozumne malé medziatómové vzdialenosti)
3. torzný člen, ktorý povoľuje len niektoré predvolené hodnoty uhlov ε

V tomto prípade je priestorová odpudivá sila a hydrofobická príťažlivá sila spolu reprezentovaná Lennard-Jonesovou funkciou, ktorá v tomto prípade vyzerá nasledujúco:

$$\sum_{|i-j|>2} \varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2H_{ij} \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{5.1}$$

Funkcia zahŕňa prírastky potenciálnej energie všetkých dvojíc oddelených viac ako o 2 v primárnom reťazci. ε_{ij} a σ_{ij} sú konštanty, ktoré vyplývajú zo vzťahov medzi jednotlivými zložkami. r_{ij} reprezentuje Euklidovskú vzdialenosť medzi guličkou i a j. Konštanta $H_{ij} = 1$ ak obe zložky sú typu H, a teda v tomto prípade sa do potenciálnej energie počítajú odpudivé sily aj príťažlivé sily. V druhom prípade , $H_{ij} = 0$ ak zložky i a j sú typu H-P, P-H alebo P-P, tvorí prírastok k celkovej potenciálnej energii len zložka odpudivých síl.

Torzný člen má tvar

$$E_\varepsilon = \sum_i C_1 \left(1 + \cos(3\varepsilon_i)\right). \tag{5.2}$$

V tomto prípade teda preferujeme uhly 60°,180° a 300°, v ostatných prípadoch dochádza k penalizácii na základe konštanty C_1. Odvodenie tohto člena je fyzikálne náročne a z hľadiska problematiky ho nieje potrebné obšírne uvádzať, je ho možno nájsť v odbornej literatúre napr. v [9].

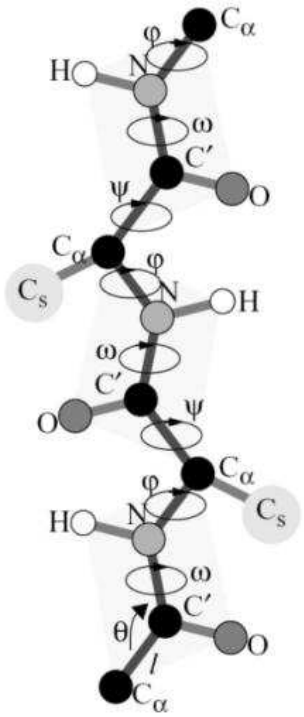
5.2 Podrobnejší polypeptidový model

Zjednodušený model v predchádzajúcej kapitole nám ponúkol náhľad na štruktúru proteínu. Tento model je síce názorný, vyhovuje na vizualizáciu problému, nepokrýva však dostatočne zložitosť polypeptidových štruktúr a nevysvetľuje všetky pôsobiace fyzikálne aspekty. Zjednodušený model je teda vhodný na otestovanie funkčnosti algoritmu, a pri jednoduchších štruktúrach dáva dobré výsledky, keďže pokrýva základné pôsobenie v polypeptidovej štruktúre.

Zložitejší model sa opiera viac o znalosti chemickej štruktúry. V reálnom proteíne je každé reziduum v primárnom reťazci charakterizované NH-C_αH-C′O časťou a jednou z dvadsiatich aminokyselín napojenou v bočnom reťazci na centrálny C_α atom. Pointou tohto modelu je, že každý postranný reťazec je klasifikovaný ako polárny alebo ako hydrofóbny a je reprezentovaný len jedným „virtuálnym“ atómom C_S. Informácie o vnútornej štruktúre sú reprezentované premennými (viď obr.5.2):

1. l -dĺžka väzby medzi dvoma atómmi v reťazci za sebou
2. θ - väzbový uhol
3. φ - väzbový klinový uhol, ktorý určuje pozíciu C′ vzhľadom k predchádzajúcim za sebou idúcimi atómmi C′-N-C_α
4. ψ - väzbový klinový uhol, ktorý určuje pozíciu N vzhľadom k predchádzajúcim za sebou idúcimi atómmi N-C_α-C′
5. ω - väzbový klinový uhol, ktorý určuje pozíciu C_α vzhľadom k predchádzajúcim za sebou idúcimi atómmi C_α- C′-N

Pre n reziduí dostávame $9n-6$ parametrov, avšak nie sú všetky navzájom nezávislé. Napr. dĺžka väzby je známa v prípade N-C′ ($1{,}32.10^{-10}$ m), C′-C_α ($1{,}53.10^{-10}$ m), C_α-N ($1{,}47.10^{-10}$ m), väzbový uhol θ je definovaný pri



Obrázok 5.2: Podrobnejší polypeptidový model, zdroj [5]

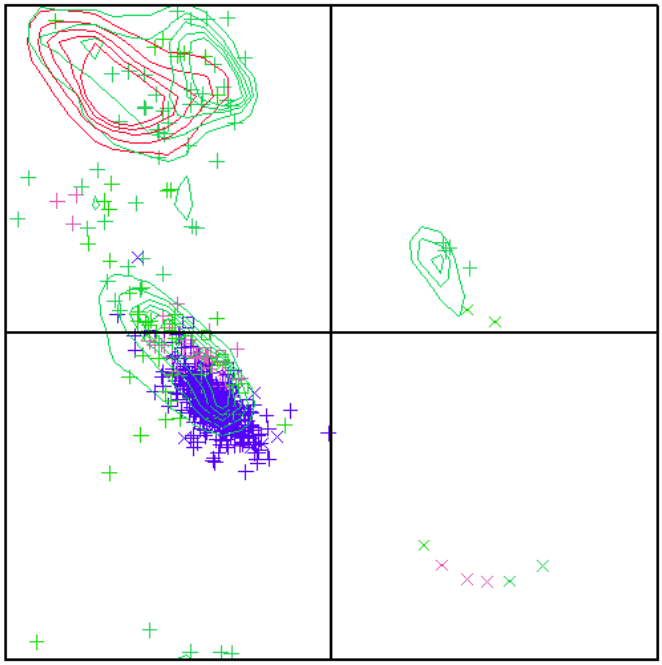
N-C_α-C′(110°), C_α-C′-N (114°), C′-N-C_α (123°)a pod. Podľa [5], na záver po odhalení všetkých pevných parametrov, dostávame $n-1$ dvojíc väzbových klinových uhlov (φ, ψ), ktoré nám reprezentujú model.

Vráťme sa ešte k reprezentácii postranného reťazca pomocou atómu C_S, kde vďaka nej nepotrebujeme žiadne dodatočné parametre. V [6] nájdeme tabuľku 20 aminokyselín, s ich klasifikáciou na polárne a hydrofóbne, s dĺžkami väzby medzi C_α a C_S, a veľkosti väzbových uhlov v sekvencii N-C_α-C_S, i veľkosť väzbových klinových uhlov medzi C_S a N-C_α-C′.

5.2.1 Funkcia potenciálnej energie

Obsahuje 3 komponenty

1. E_p výraz zahrňujúci energiu medzi rezíduami H-H
2. E_{ex} výraz zachycujúci priestorový odpudivý efekt (odmieta akékoľvek postavenie, ktoré by povoľovalo nerozumne malé medziatómové vzdialenosti)



Obrázok 5.3: Ramachandranov graf pre hemoglobín, modrá reprezentuje závitnice, červená a zelená oblúky, čiary sú preferované oblasti, krížiky typy reziduí.

3. $E_{\varphi\psi}$ torzný člen, ktorý povoľuje len tie dvojice uhlov (φ, ψ), ktoré povoľuje Ramachandranov graf (viď. napr. [9]), ktorý nám znázorňuje možné kombinácie uhlov φ, ψ pre polypeptidy. (obr.5.3)

E_p a E_{ex} sú definované následovne

$$E_{ex} = \sum_{ij} \frac{1}{1 + \exp \frac{(d_{ij} - d_{eff})}{d_W}} \tag{5.3}$$

$$E_{hp} = \sum_{|i-j|>2} \varepsilon_{ij} f(d_{ij}) \tag{5.4}$$

$$f(d_{ij}) = \frac{C_2}{1 + \exp(\frac{d_{ij} - d_0}{d_t})}$$

V E_{ex} je d_{ij} medziatómová vzdialenosť medzi dvoma atómmi C_α alebo medzi C_S, d_W=0,1.10^{-10} m je pevne daná a určuje mieru poklesu E_{ex}, d_{eff}=3,6.10^{-10} m pre atómy C_α a 3,2.10^{-10} m pre C_S, podmieňuje stredový bod funkcie (bod, v ktorom funkcia dosiahne hodnotu polovice maxima).

V E_{hp} je d_0= 6,5.10^{-10} určuje rýchlosť poklesu a d_t=2,5.10^{-10} m podmieňuje stredový bod, ε_{ij}= -1 keď reziduá i a j sú hydrofóbne a ε_{ij}= 0 inak. C_1, C_2 sú konštanty, pre podrobnosti viď. [6].

$E_{\varphi\psi}$ by mala byť typom penalizačnej funkcie, keďže ako bolo spomenuté dvojice (φ, ψ) dosahujú len niektorých hodnôt, je potrebné nevhodné dvojice penalizovať, teda aby $E_{\varphi\psi} \cong 0$ ak (φ, ψ) nadobúda vhodných hodnôt, $E_{\varphi\psi} = b$ inak, napr. ako je uvedené v [5] si všimneme z Ramachandranovho grafu, že hodnoty sa združujú v elipsách a

$$E_{\varphi\psi} = \frac{\beta}{1 + \Sigma_{i=1}^{p} \exp(-\gamma_i q_i(\varphi, \psi))} \tag{5.5}$$

Elipsoid R_i je reprezentovaný kvadratickou funkciou $q_i(\varphi, \psi)$, pre ktorú platí $q_i(\varphi, \psi) = 0$ na hranici. Vo vnútri elipsoidu platí $q_i(\varphi, \psi) < 0$, a $q_i(\varphi, \psi) > 0$ vonku, γ_i je konštanta, ktorá udáva rýchlosť s akou sa $E(\varphi, \psi)$ blíži k 0 či b.

Je potrebné podotknúť, že pri tomto zložitom modeli sa niekedy vyjadrenie energie mierne líši. Energia sa však nepočíta ručne, ale využíva sa externý program, ktorý túto energiu počíta ECEPP (Empirical Conformational Energy Program for Peptides). Pre popis algoritmov nám teda tieto informácia viac ako postačia.

Kapitola 6

Algoritmy pre problém skladania proteínov

Počítanie trojrozmerných štruktúr proteínov zahŕňa vyriešenie dvoch závažných problémov: získanie a popis spoľahlivej funkcie potenciálnej energie, a voľba vhodnej metódy prehľadávania priestoru k identifikácii globálneho minima funkcie potenciálnej energie medzi množstvom lokálnych miním.

V tejto časti budeme využívať len metódy zahŕňajúce optimalizáciu funkcie potenciálnej energie, bez využití pomocných informácií ako napr. predpoveď sekundárnej štruktúry, homologické modelovanie, informácie o proteínových štruktúrach z databáz a pod. Úpravu algoritmu týmto smerom si ilustratívne ukážeme na algoritme v kapitole (6.8). Je však potrebné podotknúť, že tieto ďalej spomínané algoritmy dobre fungujú len na menšie proteíny a fibrilárne proteíny, a len malé množstvo z nich funguje aj pre veľké systémy ako sú globulárne proteíny.

Poznámka 3 *Pokiaľ nebude uvedené inak, budeme uvažovať, že algoritmy pracujú s podrobným polypeptidovým modelom.*

6.1 Algoritmus postupnej výstavby (Build-up algorithm)

Pre proteíny z viac ako 5 jadrami sú klasické metódy výpočtu všetkých možných usporiadaní (napr. systematické prehľadávanie) nepoužiteľné. Preto bol vyvinutý nasledujúci algoritmus. (Viac informácií napr. v [8].)

V tejto metóde ide o skrátené prehľadávanie, pričom dominantný vplyv majú krátkodobé vzájomné interakcie. Procedúra najprv nájde lokálne minimá kratších fragmentov, potom tieto krátke úseky kombinuje do jedného

dlhšieho a znovu minimalizuje energiu týchto dlhších úsekov. Vhodné minimá volíme z fragmentu s najnižšou energiou. Hodnota energie týchto miním je nižšia ako zvolená medzná energia. Množina miním jedného kratšieho úseku v kombinácií s množinou ďalšieho úseku vytvoria dlhší peptid, na ktorý je opäť aplikovaný optimalizačný algoritmus. Postupne sa teda fragmenty rozrastajú. Čím dlhší reťazec máme, tým ďalekosiahlejšie budú vzájomné interakcie. Tento proces je opakovaný, kým nieje celý reťazec nakoniec zložený zo základných zložiek.

6.1.1 Algoritmus

1. Za základný stavebný prvok tejto procedúry sú považované jednotlivé jadrá aminokyselín, z ktorých budeme konštruovať konečnú štruktúru. Za medznú energiu sa zvyčajne volí hladina 5 kcal/mol.

2. Z molekuly z n reziduami dostaneme $n - 1$ dipeptidov. Po následnej minimalizácii sú dipeptidy triedené. Vhodné sú následne použité pre konštrukciu tripeptidov.

3. Generovanie väčších fragmentov bielkovinového reťazca zahŕňa spojovanie dvoch kratších úsekov, ktoré majú jedno alebo viac spoločných jadrier, napr. tetrapeptid môže vzniknúť z dvoch tripeptidov, ktoré majú spoločné 2 jadrá a pod.. Tento proces je opakovaný kým nieje postavený celý bielkovinový reťazec.

6.1.2 Problémy

Skutočnosť, že množstvo systematických usporiadaní, ktoré musia byť energeticky minimalizované a uschované v každom kroku, exponenciálne rastie, predstavuje zásadnú zápornú stránku tohto algoritmu. Popri používaniu medznej energie je čiastočným riešením udržať len tie minimá, ktorých reťazec sa markantne odlišuje od ostatných. Tento krok značne redukuje počet usporiadaní, ktoré je potrebné v každom stupni uložiť. Na druhej strane toto vylepšenie môže viesť k problémom v pokročilých stupňoch algoritmu, pretože nie vždy sú štruktúry preferované v menších reťazcoch preferované aj v tých väčších.

6.1.3 Aplikácie

Tento algoritmus bol úspešne používaný pre cyklické peptidy, fibrilárne peptidy ako napr. kolagén [15].

S výnimkou fibrirálnych proteínov, ktorých výhodou sú symetrické vzťahy, je metóda nepoužiteľná pre bielkovinový reťazec obsahujúci viac ako 20 reziduí aminokyselín.

6.2 Metóda selfkonzistentného elektrostatického poľa (The Self Consistent Electrostatic Field Method)

Táto metóda je založená na veľkom počte experimentálných údajov (napr. [17]). Tie hovoria o tom, že prirodzený stav proteínu nastáva vtedy, keď sú optimalizované elektrostatické interakcie systému, napr. dipóly peptidovej skupiny v prirodzenom stave musia mať približne optimálnu orientáciu v elektrostatickom poli generovanom celou molekulou a okolitým roztokom. Na základe tohto uvažovania vzniká prehľadávacia metóda pomenovaná Metóda selfkonzistentného elektrostatického poľa(skratka z anglického Self-Consistent Electric Field (SCEF) method), ktorá pracuje s funkciou potenciálnej energie, v ktorej je zahrnutá aj elektrostatická energia.

Poznámka 4 *Selfkonzistencia je dosiahnutie konvergencie v postupnom aproximovaní riešenia niektorých matematických a fyzikálnych problémov. Koeficienty v rovniciach, popisujúcich daný problém, sa opravujú podľa výsledkov získaných postupným aproximovaním tak dlho, až novo zostrojené koeficienty dávajú zhodné výsledky ako koeficienty predchádzajúce.*

6.2.1 Algoritmus

1. Začína z ľubovoľného usporiadania molekuly. Minimalizuje funkciu potenciálnej energie kým nieje dosiahnuté najbližšieho lokálneho minima.

2. Pre toto usporiadanie vypočíta *elektrostatické pole* molekuly v každej CO a NH skupine a tiež v centre CO-N väzby.

3. Určí smer elektrostatického poľa s ohľadom na CO a NH dipólové momenty všetkých peptidových skupín. Generuje *diagnostické rotácie*, ktoré odpovedajú variáciám, ktoré musia byť aplikované na daný torzný uhol za účelom najlepšieho zarovnania dipólu peptidovej jednotky s ohľadom na miestne elektrické pole. Elektrostatická analýza ukáže najhoršie orientovaný dipólový moment peptidovej skupiny (napr. skupina medzi jadrom i a $i+1$). Diagnostická rotácia potom opisuje zmenu väzbových klinových uhlov ψ_i a ϕ_{i+1}, potrebnú pre zaradenie dipólov.

4. Prevedie diagnostické rotácie.

5. Použije nové usporiadanie molekuly spôsobom ako v bode 1.

   - ak nájde nové lokálne minimum, opakuje procedúru od bodu 2 pre nové lokálne minimum,
   - ak nájde rovnaké lokálne minimum, opakuje 3. krok, ale použije diagnostické rotácie pre ďalší najhorší orientovaný dipól.

6. Kroky 1-5 sa opakujú v selfkonzistentom spôsobe, až kým ďalšou aplikáciou procedúry už nedôjde k zmene usporiadania molekuly.

6.2.2 Aplikácie

V [19] bola táto procedúra testovaná na 19 jadrovom poly (L-amíne). Počiatočné usporiadanie pozostávalo so sérií čiastočných α-špirál na rôznych stupňoch. Pre tento proteín pravotočivé α-špirály zodpovedajú globálnemu minimu energetickej funkcie. V štyroch prípadoch bola táto procedúra schopná dôjsť k rovnakému výsledku vo veľmi krátkom výpočetnom čase. Algoritmus bol použitý aj na 58 jadrovom proteine, vo výsledku bol schopný dosiahnuť len zlepšenia oproti počiatočnému stavu.

6.3 Minimalizačná metóda Monte Carlo (The Monte Carlo Minimization)

Skutočnosť, že proteíny nie sú statické štruktúry, ale podliehajú rôznym fluktuáciám, bola hlavným motívom rozvoja metódy Monte Carlo. Táto metóda je postavená na stochastickom prístupe ku globálnej optimalizácií, a kombinuje metódu Metropolis Monte Carlo z kombinatorickej globálnej optimalizácie s konvenčnou lokálnou minimalizáciou.

Vychádza z predpokladu, že prirodzený stav proteínu musí byť stabilný aj pri väčších termických fluktuáciách. Ak je štruktúra určená minimalizačným procesom, a je stabilná len pre malé deformácie, je veľmi pravdepodobné, že daná štruktúra bude tepelne nestabilná a z tohto dôvodu to nemôže byť vhodným kandidátom na prirodzený stav. Z týchto úvah vyplýva, že tepelné kolísanie hrá veľkú úlohu pri výbere prirodzeného usporiadania.

Aj keď MCM metóda môže simulovať termické procesy, mohli by sme uvažovať o jeho priamom použití na polypeptidové štruktúry, čo sa však ukázalo ako neúčinné. Problém je v tom, že v každom kroku modelovania zložitého priestoru usporiadaní používa len malé množstvo premenných. Ďalším problémom sa javia rozsiahle energetické obmedzenia, čím algoritmus inklinuje k vzorkovaniu vo vnútri veľmi obmedzujúceho priestoru usporiadaní. Práve preto MCM metóda zahŕňa aj konvenčnú energetickú minimalizáciu. Tak teda táto metóda generuje Markovu prechádzku na hypermriežke všetkých diskrétnych energetických miním, s Boltzmanovými pravdepodobnosťami prechodu.

6.3.1 Algoritmus

1. V prvom rade ide o výberovú stratégiu Monte Carlo, ktorá spĺňa ergodickú podmienku, že každé lokálne minimum je prístupne z akéhokoľvek iného lokálneho minima po konečnom množstve krokov.

   Máme dané usporiadanie C_{curr}^{min} s minimálnou energiou E_{curr}^{min}. Výberová stratégia sa používa na generovanie porušenej štruktúry C_{pert}. Táto stratégia pozostáva z náhodných zmien, aplikovaných na k klinových uhlov z celkového počtu N_{dieh} potrebných pre popis molekuly. Tieto zmeny sú vygenerované s pravdepodobnosťami 2^{-k} , kde $(k = 1, 2, \ldots, N_{dieh})$.

2. Na štruktúru C_{pert} aplikuje minimalizáciu jeho funkcie potenciálnej energie, do doby, kým nedosiahne najbližšieho lokálneho minima. Konečné

usporiadanie C_{pert}^{min} má energiu E_{pert}^{min} a táto štruktúra je bez prekrývajúcich sa atómov.

3. Kritérium Metropolis rozhodne, ktoré usporiadanie, či C_{curr}^{min} alebo C_{pert}^{min} je vybraté. Pravidlo vyzerá nasledujúco: ak energetický rozdiel $\Delta E = E_{pert}^{min} - E_{curr}^{min} < 0$ alebo (keď $\Delta E > 0$) ak (T je teplotný parameter, R je konštanta) je väčší ako náhodne vygenerované číslo medzi 0 a 1, potom nové usporiadanie C_{pert}^{min} nahradí aktuálne C_{curr}^{min} , inak je C_{pert}^{min} vyradené.

Poznamenajme, že T nie je fyzikálna teplota, na ktorej si prajeme predpovedať proteínovú štruktúru. Ide o číselný parameter, ktorý dohliada nad výpočtom. V kroku 3, vlastne prijímamé novú štruktúru s pravdepodobnosťou $e^{-\frac{\Delta E}{RT}}$. Čím je teda teplota nižšia aj daná pravdepodobnosť bude nižšia. Logicky teda zvyšovanie teploty zvýši pravdepodobnosť prijatia danej štruktúry.

6.3.2 Aplikácie

MCM metóda bola úspešne použitá pri štúdiu vhodných usporiadaní pre pentapeptid Met-enkephalín [12]. Samotná metóda sa však nepoužíva na dlhšie reťazce aminokyselín.

6.4 Elektrostaticky riadená metóda Monte Carlo, EDMC metóda (The Electrostatically driven Monte Carlo Method)

Elektrostaticky riadená metóda Monte Carlo je iteračnou metódou pre hľadanie usporiadania polypeptidov zložených z viac ako 20 aminokyselín. EDMC metóda kombinuje najúčinnejšie rysy minimalizačnej Monte Carlo metódy a metódy selfkonzistentného elektrostatického poľa, pričom zaradzuje aj nové techniky, ktoré vedú k účinnejšiemu prehľadávaniu stavového priestoru.

Hľadanie globálneho energetického minima molekuly prebieha ako kvázi náhodná prechádzka. Cesta nasledovaná EDMC metódou je definovaná sekvenciou energeticky minimalizovaných usporiadaní, ktoré sa náhodne stretli pri neobmedzenom počte iteračných krokoch algoritmu. V praxi je však počet iterácií konečný a je špecifikovaný užívateľom na začiatku simulácie.

Základné predpoklady stojace za EDMC metódou sú elektrostatické interakcie a predoklad stability pri tepelnom kolísaní. Oba ovplyvňujú konečné usporiadanie, štruktúru polypeptidového reťazca. Elektrostatické iterácie by mali viesť k štruktúre, ktorá predstavuje zlepšenie distribúcie náboja, (napr. očakáva sa , že nové štruktúry budú mať nižšie elektrostatické a celkové energie), zatiaľ čo tepelné fluktuácie predstavujú zmeny vo vnútri molekuly. Tieto efekty môžu viesť k tomu, že molekula príjme štruktúru s vyššou energiou, ale zároveň povoľuje proteínu uniknúť zo stabilných lokálnych miním s relatívne vysokou energiou.

Po implementácií týchto úvah dostávame hlavnú myšlienku algoritmu. Tepelné efekty sa skombinujú s náhodnými zmenami v štruktúre molekuly, napr. náhodne zmeníme náhodne vybranú podmnožinu premenných. Následne dôjde k preskupeniu elektrostatických interakcií, ako tomu bolo v SCEF metóde, dipólové momenty polypeptidu sa snažia o najlepšie zarovnanie v miestnom elektrostatickom poli produkovanom zbytkom molekuly. Naviac do metódy EDMC boli neskôr zaradené metódy, ktoré urýchľujú prehľadávanie a optimalizovanie procesu generovania nových štruktúr.

6.4.1 Algoritmus

Za prvú štruktúru algoritmu je braný neposkladaný tvar polypeptidového reťazca, ktorému sú priradené náhodne počiatočné hodnoty premenných, ktoré popisujú molekulovú štruktúru. Energia tejto štruktúry je minimalizovaná, a nasledujúce štruktúry sú získané cez série iteracií, použitím rôznorodých techník popísaných neskôr. Iteráciou procedúry rozumieme súbor manipulácii s aktuálne prijatou štruktúrou, ktorý vedie k nahradeniu aktuálnej štruktúry novovygenerovanou štruktúrou.
Generovanie štruktúr prebieha viacerými spôsobmi v niekoľkých krokoch:

a) Elektrostatická analýza podobná tej, ktorú využíva aj SCEF algoritmus, ale rozšírená o zahrnutie stáleho dipólového momentu z polárnych postranných reťazcov, je jedna z techník , ktorú algoritmus používa pri generovaní novej štruktúry. Ako prvý krok iterácie teda prebehne elektrostatická analýza aktuálnej štruktúry. Táto analýza je používaná pre určenie zarovnania stálych dipólov s miestnym elektrostatickým poľom produkovaným celou molekulou. Výsledkom sú *diagnostické rotácie*, ktoré môžu zlepšiť zrovnanie lokálnych dipólov s elektrickým poľom. Diagnostické rotácie sú zaradené na tzv. *predpovedný zoznam* možných štrukturálnych zmien, ktorý bude použitý v ďalších krokoch pri iterácii novej štruktúry.

b) Keďže žiadna z predpovedí nemusí viesť k akceptovateľnej štruktúre, používajú sa tiež náhodné vzorkovacie metódy pre generovanie ďalších štruktúr.

    1. Procedúra, v ktorej sa pozmenia klinové uhly vybraných reziduí:
        i. Náhodný výber všetkých premenných.
        ii. Náhodný výber kostrových premenných vo vnútri špecifických regiónov $\phi - \psi$ mapy.
        iii. Výber všetkých premenných v prepočítaných nízkoenergetických štruktúr, zložených z tripeptidov zahrnutých v sekvencii.
        iv. Výber kostrových premenných kompatibilných s regulárnymi štruktúrami (α-závitnica, β-list).
    2. Náhodný výber počtu reziduí ovplyvnených zmenami, a ich pozíciou v sekvencii.

Posledná implementácia EDMC metódy zahŕňa zhlukovú analýzu štruktúr. Štruktúry odpovedajúce prijatým minimám sú rozdelené do skupín a v týchto skupinách sú zoradené na základe ich celkovej energie. Naviac každá vy-

generovaná štruktúra, dokonca aj keď je odmietnutá, je pridružená k nejakej skupine. Do skupiny sa pridá len vtedy, ak je jej energia menšia než je energia najlepšieho člena skupiny. Nízko energetické štruktúry zahrnuté v ktoromkoľvek zhluku (okrem zhluku, ktorý obsahuje aktuálne minimum) môžu byť použité v iterácii pri náhodnom generovaní štruktúr popísanom v b).

Štruktúry vygenerované akoukoľvek z dvoch vyššie popísaných metód podliehajú minimalizácii celkovej energie. Aby bola novovygenerovaná štruktúra akceptovaná, musí splňovať dve kritéria:

1. Energia vygenerovaných štruktúr odpovedá už prijatému minimu, s ktorým sme sa stretli preddefinoavný počet krát. V tomto prípade je táto štruktúra automaticky vyradená z ďalšieho uvažovania. Táto analýza dlhodobých efektov zabezpečuje, že hľadanie sa nezacyklí v súbore lokálnych miním.

2. Keď štruktúra splní podmienku 1, jej energia E_{new} je porovnaná s energiou E_{curr} aktuálnej štruktúry. Následne sa aplikuje Metropolis kritérium (viď. kapitolu Minimalizačná metóda Monte Carlo).

Ak energia novej štruktúry vyhovie obom testom, potom táto nahradí aktuálnu štruktúru a začínajú nové iterácie.

Spätné prehľadávanie (Backtrack)

Počas iterácie sa môže stať, že ani zo súboru elektrostatických predpovedí, ani z náhodne generovaných štruktúr, nevzíde prijateľná nová štruktúra. Za týchto okolností algoritmus prehlási aktuálne lokálne minimum za stabilné a spustí sa procedúra backtrack. Táto procedúra zmení región prehľadávania. Buď ide o zmenou procesu generovania alebo o zásadnu zmenu v prijímaní nových štruktúr.

Táto procedúra zahŕňa tieto kroky:

a) Nový súbor štruktúr vzniká zmenením veľkého počtu premenných súčasne. Procedúra inklinuje k výberu premenných pridružených hlavne ku kostre reťazca.

b) Teplotný parameter T, používaný pri prijímaní nových štruktúr je zdvihnutý náhle na vyššiu hodnotu alebo ostáva permanentne zvýšený.

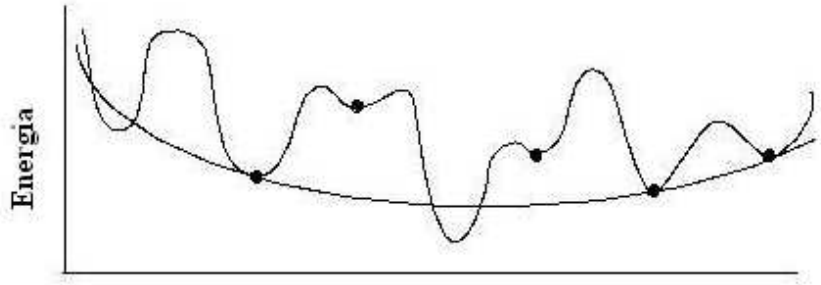
Táto procedúra postupuje kým nedôjde k akceptovaniu nejakej novej štruktúry, prípadne kým počet vygenerovaných štruktúr nedosiahne predurčenú hodnotu. V prvom prípade, sa teplotný parameter vráti k pôvodnej hodnote a generovací mechanizmus sa vráti späť do štandardného postupu popísaného vyššie. Ak nastane druhá situácia, algoritmus je zastavený na predpoklade, že prakticky je nemožné uniknúť z aktuálneho prehľadávacieho regiónu.

Tu je potrebné poznamenať, že nárast teploty behom backtrack procedúry má efekt na zvýšenie pravdepodobnosti prijatia štruktúry s energiou vyššou ako je aktuálne lokálne minimum. Daná štruktúra je však stabilnejšia. Pri tomto mechanizme sa vyhýbame uväzneniu, zacykleniu vo vysoko energetických regiónoch prehľadávacieho priestoru.

6.4.2 Aplikácie

Táto metóda sa ukázala ako vhodná pri práci s polypeptidmi o dĺžke aj viac ako 20 aminokyselín.

Konkrétne aplikácie napr. na enkephalíne [20] , oxytocíne a arginine-vasopresíne [13] ukazujú, že výsledky, ktoré dáva táto metóda, zodpovedajú globálnym energetickým minimám.



Obrázok 6.1: Príklad konvexnej interpolačnej funkcie pre CGU algoritmus

6.5 Metóda konvexného podhodnoteného odhadu (CGU (convex global underestimator) method)

Táto metóda nehľadá vrcholy v energetických plochách (ak je stav popísaný n premennými a energiou, potom v n-rozmernom priestore tvoria hodnoty energie plochu), a jej rýchlosť nezávisí na zložitosti ale závisí len na veľkosti tejto plochy. Podľa [18] je rýchlosť algoritmu pre reťazec s n stupňami voľnosti v rádoch n^4.

CGU metóda je navrhnutá tak aby interpolovala všetky známe lokálne minima konvexnou funkciou, ktorá všetky tieto minima podhodnocuje, tak aby sa od nich líšila čo najmenej v diskrétnej L_1 norme (viď.6.1).

Na túto interpoláciu môžeme použiť ľubovoľnú nezápornú lineárnu kombináciu konvexných funkcií, ale vo väčšine sa používa pre jednoduchosť konvexná kvadratická funkcia. Minimum funkcie sa používa ako predpoveď globálneho minima základnej energetickej funkcie. Na základe tohto predpovedaného minima sa vytvorí súbor polypeptidov, ktoré sú charakterizované daným minimom potenciálnej energie. Tento súbor potom slúži ako základ pre ďalší kvadratický podhodnotený odhad. Po niekoľkých opakovaniach s rozumnou istotou nájdeme globálne minimum.

Použitím podhodnocujúcej funkcie sa zložitá funkcia potenciálnej energie transformuje na jednoduchú konvexnú funkciu. Ak je teda úloha vhodná pre

použitie algoritmu CGU (teda poskytne rozumne presné predpovedi globálneho minima), dochádza k veľkej úspore času.

Prítomnosť kvadratických výrazov vo vyjadrení funkcie potenciálnej energie podporuje našu úvahu o vhodnosti použitia toho algoritmu na náš problém.

6.5.1 Popis metódy CGU pre HP model

Použijeme teda tento algoritmus na diferencovateľnú funkciu potenciálnej energie $E_{total}(\phi)$ (,ktorá pre model HP vznikne ako súčet 5.1 a 5.2), kde $\phi \in R^{n-1}$,a kde $E_{total}(\phi)$ má veľa lokálnych miním a ϕ je vektor $n-1$ uhlov medzi väzbami. Definujme $\tau = n-1$, a vypočítame množinu $k \geq 2\tau + 1$ lokálnych miním $\phi^{(j)}$, pro $j = 1, \ldots, k$.
$F(\phi)$ je konvexná kvadratická funkcia, ktorá podhodnocuje tieto lokálne minimá a normálne interpoluje $E_{total}(\phi^{(j)})$ v $2\tau + 1$ bodoch.

Týmto sú určené koeficienty vo funkcii $F(\phi)$ ako

$$\delta_j = E_{total}(\phi^{(j)}) - F(\phi^{(j)}) \geq 0 \qquad\qquad j = 1, \ldots, k \tag{6.1}$$

kde minimalizujeme $\sum_{j=1}^{k} \delta_j$. Čiže minimalizujeme rozdiel medzi $F(\phi)$ a $E_{total}(\phi)$ v diskrétnej L_1 norme, cez množinu k lokálnych miním $\phi^{(j)}$, kde $j = 1, \ldots, k$. Funkcia $F(\phi)$ je konvexná a kvadratická a najjednoduchšie ju môžeme zapísať ako

$$F(\phi) = c_0 + \sum_{i=1}^{\tau} \left(c_i \phi_i + \frac{1}{2} d_i \phi_i^2 \right) \tag{6.2}$$

c_i a d_i dostaneme z 6.1 pre každé lokálne minimum $\phi^{(j)}$. Konvexnosť tejto kvadratickej funkcie je zabezpečená podmienkou $d_i \geq 0$ pre $i = 1, \ldots, \tau$.

Dodatočne potrebujeme garantovať, že $F(\phi)$ dosiahne svojho globálneho minima v $H_\phi = \left\{ \phi_i : 0 \leq \underline{\phi_i} \leq \phi_i \leq \overline{\phi_i} \leq 2\pi \right\}$ a teda zavedieme podmienky

$$c_i + \underline{\phi_i} d_i \leq 0 \qquad\qquad c_i + \overline{\phi_i} d_i \geq 0 \qquad\qquad i = 1, \ldots, \tau \tag{6.3}$$

Tieto podmienky vedú k tomu, že $c_i \leq 0$ a $d_i \geq 0$ pre $i = 1, \ldots, \tau$. Tieto neznáme koeficienty dostaneme ako výsledok jednoduchej úlohy lineárneho programovania. Keďže konvexná kvadratická funkcia $F(\phi)$ aproximuje lokálne minimá $E_{total}(\phi)$, potom jednoducho vypočítané minimum $F_{min}(\phi)$ je teda dobrým kandidátom na globálne minimum funkcie potenciálnej energie $E_{total}(\phi)$. Detailnejší prehľad tejto úlohy lineárneho programovania je publikovaný v [27].

Konvexná kvadratická funkcia $F(\phi)$ určená hodnotami $c \in R_{\tau+1}$ a $d \in R_\tau$ nám teda poskytne globálnu aproximáciu lokálneho minima $E_{total}(\phi)$, bod globálneho minima $F_{min}(\phi)$ je v bode$(\phi_{min})_i = -c_i/d_i$, kde $i = 1, \ldots, \tau$ a $F_{min} = c_0 - \sum_{i=1}^{\tau} c_i^2/2d_i$. Ako už bolo spomenuté hodnota F_{min} je kandidátom na aproximáciu globálneho minima funkcie potenciálnej energie, a teda ϕ_{min} môže byť použitý ako počiatočný východiskový bod, v okolí ktorého budú generované ďalšie lokálne minima. Tieto lokálne minima sa pridávajú do množiny k ostatným lokálnym minimám a proces sa opakuje. Pred každou ďalšou iteráciou je potrebné zredukovať množinu H_ϕ, tak aby sme získali lepší odhad v okolí ϕ_{min}.

Ak označíme E_c je hraničná energia, potom jedným zo spôsobov ako modifikovať veľkosť H_ϕ je položiť $H_\phi = \{\phi : F(\phi) \leq E_c\}$. Pre určenie hraníc H_ϕ, uvažujeme $\{F(\phi) \leq E_c\}$, kde $F(\phi)$ spĺňa 6.2. Potom pre obmedzenie ϕ_i spĺňa podmienku

$$\sum_{i=1}^{\tau} \left(c_i\phi_i + \frac{1}{2} d_i \phi_i^2 \right) \leq E_c - c_0 \tag{6.4}$$

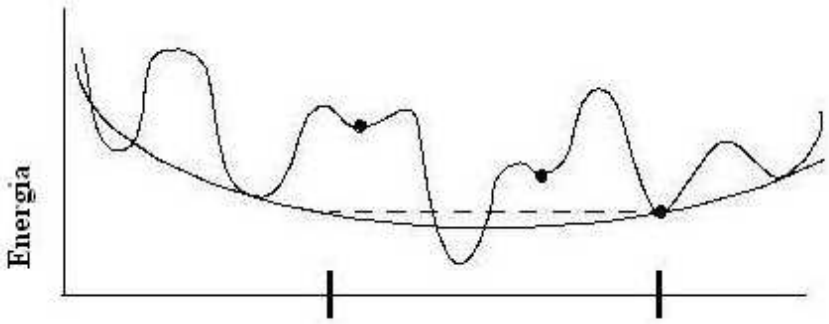
Ako už bolo spomenuté minimum $F(\phi)$ je dosiahnuté pre $\phi_i = -c_i/d_i$, kde $i = 1, \ldots, \tau$. Dosadením tejto hodnoty za každé ϕ_i mimo ϕ_k dáva

$$c_k\phi_k + \frac{1}{2} d_k \phi_k^2 \leq E_c - c_0 + \frac{1}{2} \sum_{i \neq k} \frac{c_i^2}{d_i} \equiv \beta_k \tag{6.5}$$

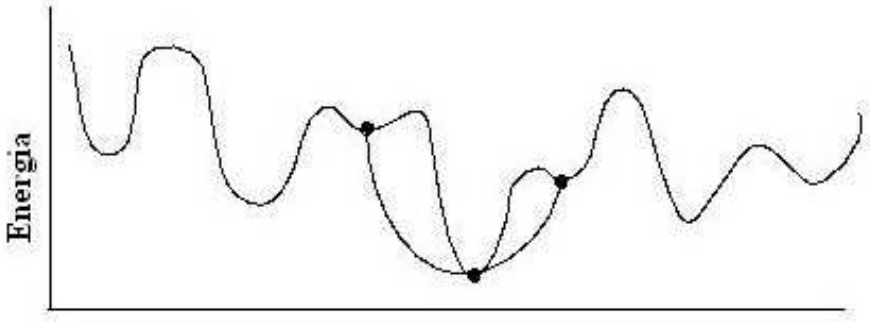
Hornú a dolnú hranicu ϕ_k, kde $k = 1, \ldots, \tau$ dostaneme ako riešenie kvadratickej rovnice $c_k\phi_k + \frac{1}{2} d_k \phi_k^2 = \beta_k$. Tieto hranice sú potom použité pri definovaní nového H_ϕ.

Jasne vyplýva, že ak zredukujeme E_c tak sa zredukuje aj H_ϕ. V každom iteračnom kroku spĺňa F_{min} podmienku $F_{min} \leq E_{total}(\phi^*)$, kde ϕ^* je najmenšia známe lokálne minimum (Obr. 6.2).

Ak v každej iterácii nájdeme jeden bod ϕ, ktorý spĺňa $E_{total}(\phi) < E_{total}(\phi^*)$, potom H_ϕ sa zmenšuje každou iteráciou, obr.6.3.



Obrázok 6.2: Definovaný H_ϕ



Obrázok 6.3: Nová funkcia na zredukovanom H_ϕ

Tento prístup orezávania H_ϕ samozrejme nezaručuje nájdenie globálneho minima. Je totiž možné, že globálne minimum môže byť vybraté z prehľadávanej množiny, k čomu môže dôjsť ak neodhalíme lokálne minimá v každej iterácii. Preto je podstatné aby počiatočná množina lokálnych miním bola dostatočne veľká, tak aby zahrnula aj skutočné globálne minimum, alebo aby globálna funcia $F(\phi)$ presne modelovala a predpokladala globálnu štruktúru $E_{total}(\phi)$. Všeobecne sa v [6] uvádza, že odhad založený na výpočetných skúsenostiach pre veľkosť počiatočnej množiny lokálnych miním je $k = 10(2\tau+1)$. Tento počet je postačujúci k zahrnutiu globálneho minima.

6.5.2 Algoritmus

1. Výpočet $k \geq 2\tau + 1$ lokálnych miním $\phi_{(j)}$, pre $j = 1, \ldots, k$ funkcie $E_{total}(\phi)$.

2. Výpočet konvexnej kvadratickej podhodnocovacej funkcie

$$F(\phi) = c_0 + \sum_{i=1}^{\tau} \left(c_i \phi_i + \frac{1}{2} d_i \phi_i^2 \right)$$

   riešením úlohy lineárneho programovania. Optimálne riešenie poskytne hodnoty c a d.

3. Výpočet bodu globálneho minima ϕ_{min}, ktorý je daný ako $(\phi_{min})_i = -c_i/d_i$, kde $i = 1, \ldots, \tau$ s odpovedajúcou funkčnou hodnotou $F_{min} = c_0 - \sum_{i=1}^{\tau} c_i^2/2d_i$.

4. Ak $\phi_{min} = \phi^*$,kde $\phi^* = argmin\{E_{total}(\phi^{(j)})\}$ je doposiaľ najlepšie nájdené lokálne minimum, potom stop a môžeme vyhlásiť ϕ^* za aproximované globálne minimum.

5. Zredukovať veľkosť H_ϕ použitím pravidla $H_\phi = \{\phi : F(\phi) \leq E_c\}$, kde $E_c = E_{total}(\phi^*)$.

6. Použiť bod ϕ_{min} ako počiatočný východiskový bod, pre generovanie ďalších lokálnych miním $\phi^{(j)}$ funkcie $E_{total}(\phi)$ na H_ϕ.

7. Vráť sa na krok 2.

6.5.3 Použitie CGU algoritmu na podrobnejší polypeptidový model

Tento algoritmus bol úspešne aplikovaný na jednoduchší model. V jednoduchom modeli boli základné komponenty NH-C_αH-C′O a prislúchajúce postranné reťazce nahradené jednoduchým virtuálnym atómom, jednotkou C. Uhol ξ nám teda nahradil, dvojicu (φ, ψ) z podrobnejšieho modelu. Pre odhalenie hydrofóbnych a polárnych efektov sme v zjednodušenom modeli klasifikovali jednotky C, buď ako polárne alebo hydrofóbne, čo je vlastne v zložitejšom modeli zakódovane v bočnom reťazci. Podľa [6], CGU algoritmus môžeme použiť nezmenený aj na tento model, použijeme funkciu $E_{total} = E_{ex} + E_{hp} + E_{\varphi\psi}$,kde $\phi \in R_\tau$ a $\tau = 2n - 2$.

6.5.4 Aplikácie

Algoritmus pre HP model bol v [6] skúšobne testovaný na niekoľkých kratších reťazcoch o dĺžke max.10 aminokyselín. HP model je v tomto prípade použitý hlavne na overnie funkčnosti algoritmu. Napriek tomu, použitím CGU na model HP a kratšie reťazce môžme dosiahnuť podobne dobré výsledky ako použitím na podrobnom modeli. Algoritmus CGU bol úspešne aplikovaný v [5] na niekoľko polypeptidových štruktúr rôznych dĺžok: Met-enkephalín (5 jadier), Bradykinín a Oxytocín (9 jadier), Mellitín (27 jadier). Ďalšie výpočty tímto algoritmom pre proteíny o dĺžkach až 30 aminokyselín nájdeme aj v [27].

6.6 Metóda simulovaného žíhania stavového priestoru (The Conformational Space Annealing (CSA) Method)

CSA algoritmus kombinuje základné kroky genetického algoritmu a algoritmu postupnej výstavby. Algoritmus prechádza v prvých krokoch celý stavový priestor, neskôr sa hľadanie špecifickým spôsobom zužuje na menšie regióny s nižšou energiou.

V úvode definujeme D_{ij} ako vzdialenosť medzi dvoma štruktúrami i a j, ktorá je zavedená ako súčet rozdielov medzi všetkými premennými klinovými uhlami. Táto vzdialenosť nám v podstate definuje podobnosť dvoch štruktúr. Zavedieme ešte medznú hranicu D_{cut}. D_{ave} označíme priemernú párovu vzdialenosť všetkých daných štruktúr.

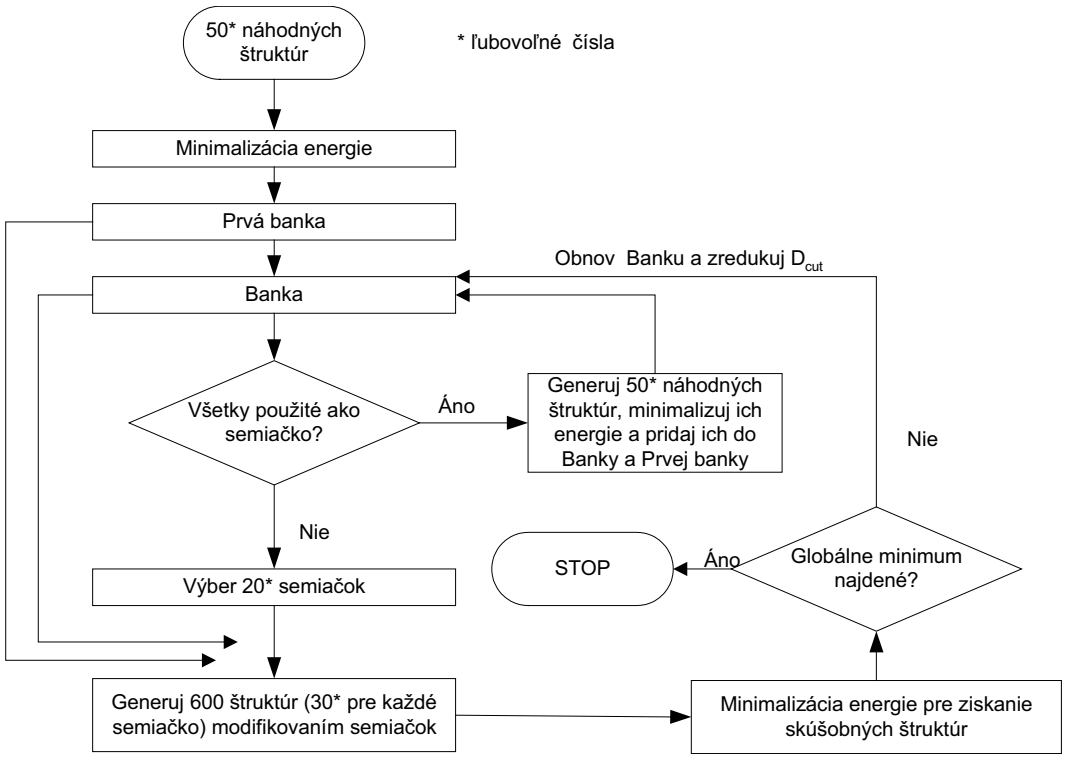
6.6.1 Algoritmus

Popis algoritmu je na obrázku 6.4. Ako v genetickom algoritme, aj CSA začína s preddefinovaným počtom náhodne vygenerovaných štruktúr, zvyčajne okolo 50. Následne sú tieto štruktúry energeticky minimalizované. Túto skupinu štruktúr nazývame *prvá banka*. Na začiatku sú štruktúry v banke rozptýlene a reprezentujú celý stavový priestor. Prvá banka slúži v ďalších krokoch ako schránka náhodných bodových mutácií klinových uhlov. Poznamenajme, že prvá banka ostáva nezmenená po celý zbytok algoritmu. Prvá banka sa následne skopíruje do *aktuálnej banky*. V ďalšom kroku je niekoľko (typicky 20) vzájomne nepodobných štruktúr, vybratých z banky. Tieto nazývame *semiačka*. Do výberu z banky nezahŕňame štruktúry, ktoré už boli v predchádzajúcich krokoch algoritmu použité ako semiačka.

Výber semiačok

Pri výbere semiačok je podstatné nevyberať štruktúry, ktoré sa nachádzajú blízko seba v stavovom priestore. Ak máme vybrať n semiačok, môžu nastať dve situácie.

- V prvom prípade je $n < m$, kde m je počet štruktúr, ktoré môžu byť použité ako semiačka (Na začiatku je m rovné počtu štruktúr v banke,



Obrázok 6.4: CSA algoritmus

zvyčajne 50). Prvé semiačko s_1 je vybrané náhodne z m bankových štruktúr, pričom v banke ostáva $m-1$ použiteľných štruktúr. Ďalej vypočítame vzdialenosti medzi semiačkom s_1 a $m-1$ bankovými štruktúrami, pričom ďalej uvažujeme iba štruktúry, ktorých vzdialenosť od s_1 je väčšia ako priemerná vzdialenosť s_1 od použiteľných bankových štruktúr. Z týchto štruktúr vyberieme s_2 ako štruktúru s najnižšou energiou, pričom ponechávame $m-2$ nepoužitých štruktúr. Ďalej opäť vypočítame vzdialenosti medzi (s_1, s_2) a $m-2$ prvkami, pričom vzdialenosť medzi (s_1, s_2) a prvkom i definujeme ako minimum vzdialenosti $D_{s_1 i}$ a $D_{s_2 i}$. Prvok s_3 je vybratý ako prvok z minimálnou energiou spomedzi štruktúr, ktorých vzdialenosť od (s_1, s_2) je viac ako priemerná vzdialenosť. Táto procedúra je opakovaná kým nevyberieme všetkých n semiačok.

- V druhom prípade ak $n > m$. Vyberieme všetkých m bankových štruktúr ako semiačka a ďalšie vyberieme náhodne zo všetkých bankových

štruktúr. Ďalšie semiačka sú vybraté ako v prvom prípade, pričom vždy zvažujeme všetky bankové štruktúry.

Skušobné štruktúry

Po výbere semiačok je každá takáto štruktúra pozmenená zmenou istého počtu premenných (od jednej premennej až po tretinu všetkých premenných). Nové premenné nie sú volené náhodne, ale sú vybraté z jednej zo zbývajúcich bankových štruktúr. Poznamenajme, že táto procedúra využíva aktuálnu banku, nie prvú banku, pretože jej zámerom je simulovať práve prechody medzi členmi.

Po následnej minimalizácii dostávame takzvané *skušobné štruktúry.* Zvyčajne sa pre každé semiačko generuje asi 30 skušobných štruktúr. Táto časť algoritmu je najviac časovo a výpočetne náročná, preto sa k jej prevedeniu často používa paralelné počítanie.

Pre každé skušobné usporiadanie α, vyberieme najbližšiu, najpodobnejšiu bankovú štruktúru A . Ak $D_{\alpha A} < D_{cut}$, potom α je podobná A. Ak má α aj nižšiu energiu ako A, v tom prípade α nahradí A v banke. Ak α nie je podobná A, ale jej energia je nižšia ako energia bankovej štruktúry B, ktorá má v banke najvyššiu energiu, potom α nahradí B v banke. Ak nenastane žiadna z predchádzajúcich dvoch možností, tak štruktúru α zamietneme.

Následne prebehne zužovanie prehľadávania a teda zmena nastavenia D_{cut} na vyššiu hodnotu (obvykle $D_{cut} = \frac{1}{2}D_{ave}$). Takto postupne redukujeme prehľadávaný stavový priestor.

Jedno kolo procedúry je kompletné, keď už nieje možné vybrať ďalšie semiačko, teda všetky bankové štruktúry už boli ako semiačka použité. Je potrebné byť dôsledný pri výbere semiačok z banky, aby každá štruktúra bola vybratá len raz. Príležitostne je potrebné zväčšiť veľkosť prvej banky pridaním niekoľkých nových náhodných štruktúr. Ide hlavne o situáciu pri ktorej síce dosiahneme medzného počtu iterácií, ale bez nájdenia globálneho minima. Kroky procedúry sa opakujú preddefinovaný počet krát.

6.6.2 Aplikácie

CSA metóda bola použitá pri počítaní globálneho minima na polypeptidy z viac ako 20 reziduami a 113 rôznymi klinovým uhlami [11]. Trochu lepšie výsledky za kratší čas dáva paralelný prístup k algoritmu, ktorý nájdeme napríklad v [10]. Avšak ani táto metóda nie je aplikovateľná na rozsiahlejšie globulárne proteíny.

6.7 Optimalizačná metóda mravenčích kolónií (Ant Colony Optimization (ACO))

Metóda mravenčích kolónií je populačná stochastická metóda vyvinutá pre riešenie širokého okruhu kombinatorických optimalizačných úloh. Algoritmus je odvodený (ako už z názvu vyplýva) zo sledovania chovania mravenčích kolónií. Skutoční mravci pri putovaní za jedlom zanechávajú feromónovú stopu, cestu, ktorú nasledujú. Takto akumulované množstvo feromónov slúži ako distribuovaná pamäť, ktorá je zdieľaná ostatnými mravcami. Izolovaný mravec sa pohybuje námatkovo, pokiaľ sa stretne s predtým položenou feromónovou cestou, zvyčajne ju nasleduje a pridá ďalšie množstvo feromónu. Čím viac mravcov prejde v nejakom čase po takej ceste, tým viac feromónov cesta obsahuje a tým je cesta atraktívnejšia. Postupom času však feromónové stopy vyprchávajú a tak sa stávajú menej atraktívne.

Základnou ideou algoritmu je teda použitie mechanizmu kladnej spätnej väzby, založenom na analógií s mravčími kolóniami. Z výpočetného hľadiska je ACO metóda iteratívna, metóda výstavbová, prehľadávacia metóda, v ktorej „obyvateľstvo jednoduchých agentov“ (mravcov) opätovne konštruuje možné riešenia daného problému. Tento stavebný proces je pravdepodobnostne riadený heuristickou informáciou o daných problémových stavoch, a zdieľanou pamäťou obsahujúcou skúsenosti mravcov z predchádzajúcich iterácií.

V tomto prípade pracujeme s HP modelom v trojrozmernej mriežke a táto metóda slúži na nájdenie optimálnej štruktúry pre danú sekvenciu.

6.7.1 Algoritmus

Táto metóda využíva kolóniu umelých mravcov, ktorí sa chovajú ako agenti v matematickom priestore, kde majú hľadať a posilniť cesty (riešenia) za účelom nájdenia optimálneho riešenia. Problém je reprezentovaný grafom a mravce sa prechádzajú po grafe a konštruujú nové riešenia.

Po inicializácií virtuálnymi ferómonmi, sa konštruujú možné riešenia a feromónové cesty sú v každom kroku aktualizované. V každom kroku sa teda vypočíta súbor možných pohybov a vyberú sa najlepšie z nich (na základe nejakých pravdepodobnostných pravidiel) a pokračuje sa v ceste. Mravce,

ktoré pracujú dobre, ovplyvňujú výskum mravcov v ďalších iteráciach, pretože tí, ako už bolo spomenuté, využívajú feromónovú cestu ako sprievodcu pri hľadaní. Pravdepodobnosť prechodu je založená na heuristickej informácií a feromónovom stupni cesty. Vyššia hodnota feromónu a heuristickej informácie vedie k výberu pohybu v danom smere.

Na začiatku je počiatočná feromónová úroveň nastavená na malú pozitívnu hodnotu konštanty τ_0 a postupne túto hodnotu mravce aktualizujú po dokončení stavebného stupňa.

Aktualizácia prebieha v dvoch stupňoch:

- lokálna aktualizácia

- globálna aktualizácia

Lokálna aktualizácia

Kým mravce budujú riešenia, zároveň lokálne aktualizujú feromónovú úroveň navštívených miest použitím lokálného pravidla:

$$\tau_{ij} \leftarrow (1 - \rho)\tau_{ij} + \rho\tau_0 \tag{6.6}$$

Kde τ_{ij} je množstvo feromónov na oblúku (i, j) na mriežke, ρ je trvácnosť cesty, a výraz $(1-\rho)$ môžeme interpretovať ako vyparovanie feromónu. Použitím tohto pravidla, mravce hľadajú v širokom okolí najlepšieho predchádzajúceho riešenia. Ako je ukázané vo formule 6.6, feromónová úroveň na cestách vysoko závisí na parametri vyparovania ρ. Jeho následkom je redukovaná úroveň feromónov a tým sa zmenšuje šanca, že ďalšie mravce vyberú rovnaké riešenie, následkom toho dochádza k rôznorodosti možných riešení.

Do algoritmu je tak implementované časové merítko, ktoré však nesmie byť príliš dlhé, inak by riešenie mohlo ostať uväznené v lokálnom minime. Nesmie však byť ani príliš krátke, inak by nedošlo k využitiu efektu spolupráce. Keď teda mravec navštívi hranu, aplikácia pravidla lokálnej aktualizácie zapríčiní zníženie hladiny feromónu v hrane, čo má vplyv na vytváranie efektu menšej atraktivity navštívených hrán, čo nepriamo vedie k povzbudeniu záujmu o doposiaľ nenavštívené hrany. Keď mravce skúmajú rôzne cesty, je tu väčšia pravdepodobnosť, že jeden z nich nájde lepšie riešenie, než v prípade, že všetci konvergujú k rovnakej ceste. Touto cestou mravce lepšie využívajú feromónovú informáciu: bez lokálnej aktualizácie by všetky mravce mohli hľadať v tesnej blízkosti najlepšej predchádzajúcej trasy.

Globálna aktualizácia

Ak mravce dokončili výstavbu riešení, je feromónová úroveň aktualizovaná použitím globálnej aktualizácie (6.7). Je použitá len na cesty patriace najlepšiemu riešeniu, tzn., že len mravec, ktorý vygeneroval najlepšiu cestu od začiatku môže globálne aktualizovať koncentráciu feromónu.

$$\tau ij \leftarrow (1-\rho)\tau_{ij} + \rho\Delta\tau_{ij} \tag{6.7}$$

$$\Delta\tau_{ij} = \begin{cases} -E_{gb}, & (i,j) \in Naj \\ 0, & inak \end{cases}$$

E_{gb} je energia najlepšej štruktúry v tomto kroku. Naj značí najlepšiu štruktúru v danom kroku. Toto globálne pravidlo má za účel poskytnúť väčšie množstvo feromónu na cestách najlepšieho riešenia, tým zintenzívni hľadanie v okolí tohto riešenia.

V mriežke existuje 6 rôznych pozícií pre každú aminokyselinu. Ak je štruktúra rotačne invariantná, pozície prvých dvoch aminokyselín môžu byť zafixované bez straty obecnosti. Počas stavebnej fázy mravce skladajú proteín z ľavého konca pridávaním aminokyselín v závislosti od hodnoty feromónovej matice (, ktorá reprezentuje predchádzajúce vyhľadávajúce skúsenosti) a od heuristickej informácie.

Pravdepodobnosť výberu pozície ďalšej aminokyseliny je daná:

$$P_{ij} = \begin{cases} \frac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum_{s\in I}\tau_{is}^{\alpha}\eta_{is}^{\beta}}, & j \in I \\ \\ 0, & inak \end{cases}$$

Kde τ_{ij} je množstvo feromónu uložené každým mravcom na ceste (i,j), α je kontrolný parameter intenzity, η_{ij} je heuristická informácia ekvivalentná počtu nových H-H kontaktov, ak vyberieme j-tú pozíciu, β je heuristický parameter, I je množina voľných susedných pozícií.

Čím vyššie sú hodnoty τ_{ij} a η_{ij}, tým výhodnejšie je vybrať aminokyselinu na j-tej pozícii. Ak ďalšia aminokyselina je polárna, pravdepodobnosť je $P_{ij} = 0$. V tomto prípade je nasledujúca pozícia zvolená náhodne spomedzi povolených pozícií. Pokiaľ je množina I prázdna, mravec spraví niekoľko krokov späť a ďalej pokračuje v budovaní riešenia.

6.7.2 Aplikácie

Algoritmus bol testovaný na štandardnom testovacom reťazci o dĺžke 48 aminokyselín, pri použití 5 mravcov [22]. V porovnaní z niekoľkými metódami pracujúcimi na základe hydrofobických jadier ako napr. hydrophobic zipper, je najlepšie riešenie tejto metódy lepšie ako najlepšie riešenia iných metód.

6.8 Stochastická - odchýlková metóda globálnej optimalizácie (Stochastic-pertrubation global optimization method)

Táto metóda pozostáva z dvoch častí. Prvá *inicializačná fáza* generuje súbor usporiadaní, ktoré lokálne minimalizujú energetickú funkciu a majú rozumne nízku hodnotu energie. Druhá fáza obsahuje väčšinu výpočetnej sily, je to tzv. *vylepšovacia fáza*, v ktorej nachádzame nové nižšie lokálne minimá smerom z aktuálne najnižšieho lokálneho minima. Kľúčovým krokom tejto fázy je riešenie globálnej optimalizácie s malý počtom parametrov (typicky 6-9 parametrov) stochastickou metódou.

6.8.1 Prvá fáza

Základný prístup

V tejto fáze teda generujeme základne vhodné vstupné usporiadanie. Na rozdiel od niektorých základných typov generovania, nie je v reťazci možné hýbať atómmi bez ovplyvnenia väzbových dĺžok a uhlov napojených na susedné atómy, teda generovanie musí prebehnúť inak. Reťazec môžeme zložiť z klinových uhlov. Zvolíme počiatočný a postupujeme volením ďalších pozdĺž celého reťazca. Každý klinový uhol je fixovaný viackrát a vyberie sa hodnota toho, ktorého čiastočný reťazec má najmenšiu energiu. Táto fáza dáva priestor pre zlepšenia, napríklad využitím známych usporiadaní.

Pokročilejší postup

V prvej fáze sa môže využiť externej informácie z proteínových databáz, ktoré sú bežne dostupné na serveroch. Vďaka týmto informáciám môžeme získať predpoveď sekundárnej štruktúry s vyššou pravdepodobnosťou.

V tejto fáze sa buduje počiatočné usporiadanie a predpovedá sa sekundárna štruktúra. Ako sme už spomenuli využíva sa externých informácií. Server predpovedá, či každá aminokyselina je prítomná v α-závite, či β-liste, na základe znalosti z ďalších známych proteínov. Používa sa aj penalizačná funkcia, ktorá slúži k upraveniu štruktúry smerujúcemu k predpovedanej sekundárnej štruktúre. Penalizačné funkcie podporujúce α, β štruktúry nájdeme napr. v [7].
Táto procedúra minimalizuje sumu potenciálnej energie a penalizačnej funkcie za účelom dostať sa do zhody s predpovedanou sekundárnou štruktúrou.

Najlepšie štruktúry vygenerované touto procedúrou sú vybrané ako počiatočné body pre celkovú lokálnu minimalizáciu, niekoľko najlepších miním prechádza do druhej fázy algoritmu.

6.8.2 Druhá fáza

V druhej fáze vylepšujeme špecifickým spôsobom danú štruktúru, použitím globálnej stochastickej metódy na malú podmnožinu parametrov, a lokálnej optimalizácie na úplnú štruktúru.

Základným krokom je výber takej podmnožiny parametrov, ktoré zmenou ich hodnoty môžu viesť k podstatnému vylepšeniu konečnej štruktúry. Dostávame zoznam všetkých miním, získaných lokálnou minimalizáciou, a po istom počte iterácii, ich rozdelíme do skupín na základe odmocniny smerodatnej odchýlky (pairwise root mean square deviation, RMSD) a v rámci skupín sa zoradia podľa hodnoty energie. RMSD dvoch štruktúr vyjadruje podobnosť štruktúr. Pokiaľ nebolo dosiahnuté stop kritéria popísaného nižšie, proces sa opakuje s novým zoznamom miním, získaných výberom prvkov z najnižšou energiou z každej skupiny.

Základná myšlienka tejto fázy je výber vhodnej štruktúry zo zoznamu lokálnych miním, následne výber malej podmnožiny jej premenných, v závere lokálna minimalizácia (za prítomnosti všetkých premenných) z najlepších výsledných štruktúr.

Pri výbere vhodného usporiadania môžeme využiť model stromového vyhľadávania. Strom je zostavený z počiatočných miním a všetkých ďalších miním vygenerovaných doposiaľ použitím globálnej optimalizácie na malom podpriestore, a použitím lokálnej minimalizácie na celý priestor. V počiatočných iteráciach vyberieme strom s najmenším množstvom práce vykonanej doposiaľ na jeho členoch. Práve v tomto strome vyberieme štruktúru, ktorú budeme vylepšovať, ide práve o štruktúru s najnižšou energiou.

Po vybratí vhodnej konfigurácie vyberieme aj podmnožinu jej premenných pre aplikovanie globálnej optimalizácie. Táto podmnožina pozostáva z malého množstva klinových uhlov proteínu náhodne vybraného, pričom ostatné nevybraté ostávajú dočasne pevné. Výber uhlov cez ktoré budeme optimalizovať je veľmi dôležitý pre úspešnosť tejto metódy. Vybraním malého počtu uhlov, ktoré sú rozptýlené pozdĺž celého proteínového reťazca docielime najväčšiu variabilitu v terciálnej štruktúre. Z výpočetného hľadiska

je však táto možnosť rovnako náročná ako celková optimalizácia. Opačne, nevhodným výberom dostaneme menšiu možnosť zmeny v celkovej štruktúre v priebehu tejto osekanej optimalizácie. Preto je dôležitý správny výber uhlov. My pri výbere použijeme energetickú funkciu, ktorá nám bude signalizovať, ktoré parametre majú najlepší potenciál k zníženiu hodnoty energetickej funkcie. Tento algoritmus k tomuto využíva dve rôzne metódy, obe založené na interakčnej energii medzi všetkými prvkami vľavo od daného klinového uhlu a medzi všetkými napravo od uhlu. Navzájom sa líšia v normalizácii interakčnej energií. Prvá metóda vypočíta interakčnú energiu pre každý klinový uhol a štandardizuje ju súčinom počtu atómov naľavo a počtom atómov napravo. Následne je vybraný špecifický počet (okolo 5) uhlov s najvyššou normovanou interakčnou energiou. Druhá metóda štandardizuje energiu maximom počtu atómov vľavo a vpravo. Prvá metóda inklinuje k výberu uhlov skôr zo strednej časti usporiadania, kým druhá metóda vyberá uhly skôr od konca. Práve preto sa prvá metóda zdá byť vhodnejšia v skorších stupňoch algoritmu, keď je počítanie ešte ďaleko do globálneho minima, zatiaľ čo druhá sa používa k „doladeniu“ , už po zoptimalizovaní stredovej časti, pri optimalizácii krajných častí. V tejto časti je vhodný priestor na heuristiku, napr. opierajúcu sa o vlastnosti a prítomnosť väzieb ako napr. v [2].

Lokálna optimalizácia je prevedená algoritmom BFGS (Broyden-Fletcher-Goldfarb-Shanno method), čo je algoritmus používaný na riešenie nelineárnych optimalizačných úloh. Je to metóda odvodená z Newtonovej metódy, táto technika hľadá stacionárny bod funkcie, kde gradient je 0. Táto metóda nieje vhodná pre veľké prehľadávacie priestory. Táto optimalizácia na obmedzenom menšom priestore nám dovoľuje preskúmať usporiadania s inými tvarmi než tie, s ktorými sme začínali v prvej fáze. Niekoľko z týchto usporiadaní s najnižšími energetickými hodnotami je vybraných pre lokálnu minimalizáciu na celom priestore premenných. Táto lokálna optimalizácia už markantne nemení štruktúru, skôr dochádza k lokálnemu zjemňovaniu štruktúr, ktoré sú však dôležité pre konečnú podobu proteínu. Lokálna minimalizácia sa robí v karteziánskych súradniciach, množstvo premenných je tak veľké, že sa k tomuto účelu použije pamäťovo obmedzená kvazi-Newtonova metóda L-BFGS (Limited memory BFGS method). Novo nájdené minimum je pridané k ostatným a celá fáza je opakovaná presný počet opakovaní.

Pravidelne sa v druhej fáze minimá zoradia a rozdelia, každá skupina je definovaná tak, že členy majú hodnotu RMSD=5.100×10^{-9} mm od e-

nergeticky najnižšieho usporiadania v tomto zhluku. Počet skupín, zhlukov signalizuje počet rôznorodých štruktúr, ktoré existujú na danom stupni. Pokiaľ energia globálneho minima už neklesá, potom algoritmus je považovaný za konvergentný. Ak sa teda energia globálneho minima znižuje, vykonáme ďalšie iterácie. V tomto prípade je za nový zoznam miním stanovený zoznam najnižších energetických miním z každého zhluku.

6.8.3 Algoritmus bez využitia externých zdrojov

1. *Počiatočné generovanie*

   a) Proteínové zostavenie vzorku
   Vybuduje vzorok usporiadania z jedného konca proteínu k druhému, generovaním každého klinového uhlu. Aktuálnu hodnotu uhlu zafixujeme a nasledujúci vyberá tak, aby energetická hodnota doposiaľ vygenerovaného proteínu bola čo najnižšia.

   b) Výber východzích bodov
   Vyberie vhodnú podmnožinu vzorkov z kroku a) ako východzie body pre lokálnu minimalizáciu.

   c) Lokálna minimalizácia na celom priestore premenných
   Z každého bodu z 1b) prevedie lokálnu minimalizáciu. Uschová tieto minimá pre fázu 2.

2. *Vylepšenie lokálnych miním*
   Prevedie isté množstvo iterácií.

   a) Výber usporiadania a podmnožiny parametrov
   Zo zoznamu lokálnych miním z prvej fázy vyberie adepta na vylepšenie. Potom vyberie podmnožinu paramtetrov z tejto konfigurácie a v ďalšom kroku optimalizuje.

   b) Globálna optimalizácia podproblému
   Použije globálny optimalizačný algoritmus na energetickú funkciu len s vybranými parametrami ako premennými.

   c) Lokálna minimalizácia
   Použije lokálnu minimalizačnú procedúru so všetkými premennými s najnižším energetickým usporiadaním, ktoré vyplynulo z predchádzajúceho kroku. Nové lokálne minimá pridá do zoznamu miním.

d) Zoskupenie lokálných miním a test konvergencie
Zoskupí všetky lokálne minimá na základe odmocniny smerodatnej odchýlky. Pokiaľ nebolo dosiahnuté stop kritéria, opakuje všetky kroky druhej fázy na novom zozname miním, ktorý obsahuje minimá z najnižšou energiou z každej skupiny.

6.8.4 Algoritmus s využitím externých zdrojov

1. *Prvá fáza*
   Generuje počiatočné usporiadanie

   a) Vzorkovanie v plnej doméne
   Generuje parametre z nejakých vzorových usporiadaní

   b) Skreslená lokálna minimalizácia
   Využitím informácii z databáz vytvorí podmienky pre predpovedanú sekundárnu štruktúru.

   c) Neskreslená lokálna minimalizácia
   Prevedie lokálnu minimalizáciu každého skresleného minimalizovania použitím neskreslenej energetickej funkcie. Uschová tieto minima pre krok druhú fázu.

2. *Vylepšenie lokálnych miním*
   Prevedie isté množstvo iterácií.

   a) Výber usporiadania a podmnožiny parametrov
   Zo zoznamu lokálnych miním z prvej fázy vyberie adepta na vylepšenie. Potom vyberie podmnožinu paramtetrov z tejto konfigurácie a v ďalšom kroku optimalizuje.

   b) Globálna optimalizácia podproblému
   Použije globálny optimalizačný algoritmus na energetickú funkciu len s vybranými parametrami ako premennými.

   c) Lokálna minimalizácia
   Použije lokálnu minimalizačnú procedúru so všetkými premennými s najnižším energetickým usporiadaním, ktoré vyplynulo z predchádzajúceho kroku. Nové lokálne minimá pridá do zoznamu miním.

   d) Zoskupenie lokálných miním a test konvergencie
   Zoskupí všetky lokálne minimá na základe odmocniny smerodatnej odchýlky. Pokiaľ nebolo dosiahnuté stop kritéria, opakuje všetky kroky druhej fázy na novom zozname miním, ktorý obsahuje minimá z najnižšou energiou z každej skupiny.

6.8.5 Aplikácie

Testovanie algoritmu bez využitia externých zdrojov prebiehalo na proteíne polyalaníne, ktorý sa používa práve k tomuto účelu, práve kôli svojej jednoduchosti. Je zložený len z jedného typy aminokyselín. V [2] je táto medtóda aplikovaná na polyalanín o dĺžke 20, 30, 40 aminokyselín. V každom prípade bolo nájdene globálne minimum a prirozdený stav proteínu mal tvar pravotočivej α-závitnice.

Použitie algoritmu, ktorý využíva aj externej informácie nájdeme napr. v [7]. Tento prístup sa ukázal vhodný najmä v prípadoch, kde daná sekvencia aminokyselín neodpovedala presnému tvaru v databáze.

Kapitola 7

Porovnanie algoritmov

V tejto kapoitole by sme sa chceli venovať stručnému porovnaniu uvedených algoritmov. Keďže v bežných podmienkach nieje možné dané algoritmy prakticky vyskúšať na probléme skladania proteínov (výpočty väčšinou prebiehajú paralelným spôsobom na niekoľkých desiatkach procesorov), porovnáme ich na základe publikovaných výsledkov.

Prvým spomínaným algoritmom je algoritmus postupnej výstavby. Ako sme už uviedli, skutočnosť, že množstvo systematických usporiadaní, ktoré musia byť energeticky minimalizované a uschované v každom kroku, exponenciálne rastie, predstavuje zásadnú zápornú stránku tohto algoritmu. Je tu síce priestor pre isté vylepšenia, avšak nie do takej miery aby sme mohli tento algoritmus použiť na rozsiahlejšie štruktúry.

Metóda selfkonzistentného elektrostatického poľa využíva ďalších fyzikálnych prístupov k problému skladania proteínov, a teda charakteristickým spôsobom prehľadáva stavový priestor. Ako však ukazujú výsledky, podobne ako ďalšia samostatne použitá minimalizačná Monte Carlo metóda, opäť nieje vhodná na aplikáciu na dlhšie polypeptidové reťazce. Metóda selfkonzistentného elektrostatického poľa však poukazuje na jeden zo základných problémov riešenia proteínovej štruktúry: aproximovanú funkciu potenciálnej energie. Ktorýkoľvek z používaných tvarov energetickej funkcie totiž nepopisuje úplne všetky aspekty ovplyvňujúce jej hodnotu. Preto jedna z ciest k vylepšovaniu algoritmov použiteľných na problém skladania proteínov je práve úprava energetickej funkcie, tak by čo najlepšie popisovala realitu.

Elektrostaticky riadená metóda Monte Carlo, dáva dokopy najúčinnejšie kroky predchádzajúcich dvoch metód, čím dostávame metódu , ktorá úspešne pracuje už aj s reťazcami o dĺžke 20 aminokyselín. V tomto smere sa opäť potvrdzuje domnienka, že správnym zachytením fyzikálnych javov v proteínoch a ich vhodnou aplikáciou do algoritmu dostávame lepšie výsledky.

Ďalšou spomenutou metódou je metóda konvexného podhodnoteného odhadu, ktorá pracuje na podstatne iných princípoch ako iné algoritmy. Klasické algoritmy globálnej optimalizácie prehľadávajú stavový priestor, po nájdení lokálneho minima, im zaberie dlhý čas návrat k hľadaniu globálneho minima. Pokiaľ algoritmus nieje schopný posunúť sa z lokálneho minima (čo sa často stáva pri klasickom prehľadávaní stavového priestoru), vravíme, že uviazol v kinetickej pasci. Zložitosť CGU algoritmu nezávisí na zložitosti a tvare energetickej plochy ale len na jej veľkosti, a teda závisí na dĺžke reťazca. Táto metóda je dobrým ekvivalentom klasických metód a dáva porovnateľné výsledky na kratších reťazcoch, dokonca za kratší čas.

Metóda simulovaného žíhania stavového priestoru patrí ku klasickým prístupom globálnej optimalizácie. Nevýhodou tohto algoritmu ostáva časová náročnosť.

Zástupcom so skupiny genetických algoritmov v našom prípade je metóda mravenčích kolónií. Jej nevýhodou v porovnaní s ostatnými algoritmami ostáva jej použiteľnosť len na HP model, aj keď na tomto modeli dáva jedny z najlepších výsledkov. Vďaka zjednodušenému modelu je však tento algoritmus aplikovateľný na dlhšie polypeptidové reťazce.

Stochastická - odchýlková metóda globálnej optimalizácie, hlavne algoritmus s využitím externej informácie nám ukazuje cestu, ktorou sa v súčasnosti uberá vývoj algoritmov v tejto oblasti. Ide hlavne o kombináciu využitia už známych štruktúr, či iných znalosti o podobe proteínu a ich vhodné zahrnutie do algoritmu. V tomto smere je veľký potenciál pre skrátenie výpočetnej doby a možnosti využitia na dlhšie polypeptidové reťazce.

Kapitola 8

Záver

Cieľom tejto práce bolo poskytnúť prehľad metód globálnej optimalizácie na praktických úlohách. Ako konkrétny príklad bol zvolený problém skladania proteínov, na ktorom boli aplikované niektoré algoritmy. Jedným z vhodných riešení sa javí kombinácia matematických metód a externých informácií o už známych štruktúrach, alebo heuristika založená na špeciálnych vlastnostiach proteínu. Je však potrebné poznamenať, že bez vhodnej funkcie potenciálnej energie, ktorá reálne vystihuje stav štruktúry, nemôžme očakávať presné výsledky zodpovedajúce realite.

Autorkinou snahou bolo z veľkého množstva dostupnej literatúry získať dostatočny prehľad na priblíženie tejto problematiky čitateľovi, a popísať niektoré často využívané optimalizačné metódy pri riešení tohto zložitého problému.

Literatúra

[1] GThang N. Bui and Gnanasekaran Sundarraj. An efficient genetic algorithm for predicting protein tertiary structures in the 2d hp model. 2005.

[2] R. Byrd, E. Eskow, A. Hoek, R.B.and Chung-Shang Shao Schnabel, and Zhihong Zou. Global optimization methods for protein folding problems. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 1994.

[3] L. Collatz and W. Wetterling. *Optimization problems*. Applied Mathematical Sciences. Springer, New York, 1975.

[4] Scott E. Decatur. Protein folding in the generalized hydrophobic-polar model on the triangular lattice. 1996.

[5] K. A. Dill, A. T. Phillips, and J. B. Rosen. Molecular structure prediction by global optimization. 1996.

[6] K. A. Dill, A. T. Phillips, and J. B. Rosen. Cgu : An algorithm for molecular structure prediction. *The IMA Volumes in Mathematics and its Applications*, 94:1–21, 1997.

[7] E. Eskow, B. Bader, R. Byrd, S. Crivelli, T. Head-Gordon, V. Lamberti, and R. Schnabe. An optimization approach to the problem of protein structure prediction. 2004.

[8] K.D. Gibson and H.A. Scheraga. Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization. *J. Comput. Chem.*, 8:826–834, 1987.

[9] A.R. Leach. *Molecular Modelling, Principles and Applications*. Longman, London, 1996.

[10] J. Lee and H.A. Scheraga. Conformational space annealing by parallel computations: extensive conformational search of met-enkephalin and of the 20-residue membrane-bound portion of melittin. *Int. J. Quant. Chem.*, 75:255–265, 1999.

[11] J. Lee, H.A. Scheraga, and S. Rackovsky. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. *Biopolymers*, 46:103–115, 1998.

[12] Z. Li and H.A. Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci.*, 84:6611–6615, 1987.

[13] A. Liwo, A. Tempczyk, M.D. Ołdziej, S.and Shenderovich, V.J. Hruby, S. Talluri, J. Ciarkowski, F. Kasprzykowski, L. Łankiewicz, and Z. Grzonka. Exploration of the conformational space of oxytocin and arginine-vasopressin using the electrostatically-driven monte carlo and molecular dynamics metho. *Biopolymers*, 38:157–175, 1996.

[14] M. Maňas. *Optimalizační úlohy.* Nakladatelství technické literatury, Praha, 1979.

[15] M.H. Miller and H.A. Scheraga. Calculation of the structures of collagen models. role of interchain interactions in determining the triple-helical coiled-coil conformation. 1.poly(glycyl-prolyl-prolyl). *J. Polymer Sci.: Polymer Symposia*, 54:171–200.

[16] Panos M. Pardalos, Vladimir L. Boginski, Oleg Alexan Prokopyev, Wichai Suharitdamrong, Paul R.Carney, Wanpracha Chaovalitwongse, and Alkis Vazacopoulos. *Optimization techniques in medecine.* Essays and Surveys in Global Optimization. Springer, US, 2005.

[17] M.F. Perutz. Electrostatic effects in proteins. *Science*, 201:1187–1191, 1978.

[18] A.T. Phillips, J.B. Rosen, and K.A. Dill. *Convex global underestimation for molecular structure prediction.* From Local to Global Optimization. Kluwer Academic Publishers, US, 2001.

[19] L. Piela and H.A. Scheraga. On the multiple-minima problem in the conformational analysis of polypeptides. i. backbone degrees of freedom for a perturbed helix. *Biopolymers*, 26:33–58, 1987.

[20] D.R. Ripoll and H.A. Scheraga. The multiple-minima problem in the conformational analysis of polypeptides. iii. an electrostatically driven monte carlo method; tests on enkephalin. *J. Protein Chem.*, 8:263–287, 1989.

[21] Gareth John Rylance and Roy L. Johnston. Applications of genetic algorithms in protein folding studies. 2004.

[22] Fidanova S. 3d hp protein folding problem using ant algorithm. *BioPS'06*, pages III.19–III.26, 2006.

[23] H.B. Thompson. Calculation of cartesian coordinates and their derivatives from internal molecular coordinates. *The Journal of Chemical Physics*, 47:3407–3410, 1967.

[24] A. Tolstoy. Optimization issues in ocean acoustics. *The IMA Volumes in Mathematics and its Applications*, 92:151–172, 1997.

[25] A. Törn and A. Žilinskas. *Global optmization*. Springer-Verlag, Berlin Heidelberg, 1989.

[26] K. Uutela, M. Hamalainen, and R. Salmelin. Global optimization in the localization of neuromagnetic sources. *Biomedical Engineering, IEEE Transactions*, 45:716–723, 1999.

[27] V.H. Walke, A. T. Phillips, and J. B. Rosen. Molecular structure determination by convex global underestimation of local energy minima. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 23:181–198, 1996.