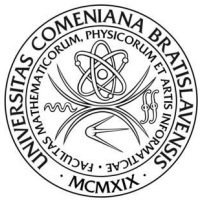
Univerzita Komenského v Bratislave

Fakulta matematiky, fyziky a informatiky



Mgr. Martin Macko

Autoreferát dizertačnej práce

ALGORITMUS PRE HĽADANIE VZDIALENÝCH FUNKČNÝCH ORTOLÓGOV A JEHO APLIKÁCIA NA OB-FOLD DOMÉNU

na získanie akademického titulu philosophiae doctor

v odbore doktorandského štúdia:

Informatika

Miesto a dátum:
Bratislava 31.7.2012

Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia na Katedre aplikovanej Informatiky, Fakulty matematiky, fyziky a informatiky Univerzity Komenského v Bratislave.

Predkladateľ: Mgr. Martin Macko
Katedra aplikovanej informatike
FMFI UK v Bratislave
Mlynská dolina, 842 48 Bratislava
(meno a priezvisko predkladateľa a adresa jeho pracoviska)

Školiteľ: Mgr. Tomáš Vinař PhD.
Katedra aplikovanej informatike
FMFI UK v Bratislave
Mlynská dolina, 842 48 Bratislava

Oponenti: .....
.....
.....
.....
.....
.....

Obhajoba dizertačnej práce sa koná ..... o ..... h
pred komisiou pre obhajobu dizertačnej práce v odbore doktorandského štúdia vymenovanou predsedom odborovej komisie .....

9.2.1 Informatika

na Fakulte matematiky, fyziky a informatiky Univerzity Komenského
Mlynská dolina
842 48 Bratislava

Predseda odborovej komisie:
.....
(meno a priezvisko s uvedením titulov a hodností
a presná adresa jeho zamestnávateľa)

Abstract

Identification of evolutionarily related proteins (orthologs) is an important step towards understanding protein function in living organisms. This problem is difficult in many cases because distantly related proteins have often too divergent amino-acid sequences, so this relation may not be identified by traditional methods based mainly on sequence similarity. In these cases search should use combination of structural and sequential features of protein.
In this thesis, we propose a new approach to the search for remote functional orthologs, where important features of protein sequence and structure are represented by a descriptor created by human expert. This descriptor can be used to search for described features in candidate protein sequences. For the problem of ortholog identification with descriptor, we develop scoring scheme which combines sequence profiles and support vector machines to evaluate alignment of the descriptor to the candidate sequence, and we develop algorithms which that the best possible alignment.
We demonstrate our approach on the example of telomere-binding OB-fold domain. Our method can distinguish between Telo_bind family members and negatives, and also identifies proteins containing a related OB-fold domain.

Keywords: protein ortholog identification, support vector machines, integer linear programming, dynamic programming

Úvod

Základom fungovania živých organizmov je využívanie genetickej informácie uloženej v chromozómoch na tvorbu rozličných proteínov. Biologická funkcia proteínu v bunke závisí nielen od sekvencie aminokyselín, z ktorej sa skladá, ale aj od tvaru štruktúry, ktorú zaujme v trojrozmernom priestore. Niektoré proteíny napriek tomu, že vystupujú v esenciálnych biologických procesoch, nemajú v rámci evolúcie výrazne zachovanú informáciu o biologickej funkcii v sekvencii aminokyselín, ale v 3D štruktúre, čo výrazne sťažuje vyhľadávanie týchto proteínov pomocou konvenčných metód založených na sekvenčnej podobnosti proteínov.

V dizertačnej práci sme popísali metódy, ktoré sa zaoberajú hľadaním konkrétnych génov v evolučne príbuzných organizmoch, rozoberali sme najčastejšie používané nástroje pri tejto analýze a rôzne prístupy, ktoré sa používajú na zlepšenie senzitivity tohto hľadania. Tieto metódy sú úspešné pre veľké množstvo génov, ale nie vždy nájdu gény pre evolučne vzdialenejšie proteíny.

Preto je cieľom dizertačnej práce vytvoriť reprezentáciu (deskriptor) štrukturálnej domény hľadaného proteínu (úsek sekvencie so špecifickými vlastnosťami) na základe zachovaných spoločných prvkov v sekvencii a štruktúre evolučne príbuzných proteínov. Ďalej je potrebné vyvinúť efektívny algoritmus pre identifikáciu proteínov s doménou popísanou týmto deskriptorom. Takýto prístup by mohol rozšíriť senzitivitu hľadania na proteíny, ktoré sa sekvenciou výraznejšie odlišujú od známych proteínov s danou štrukturálnou doménou.

Kapitola 1

Problém hľadania ortologických komplexov

Kompletná genetická informácia organizmu sa nazýva *genóm*. *Gén* je úsek genetickej sekvencie zodpovedajúci niektorej funkčnej jednotke (proteín, RNA gén a podobne). Z génov sa v bunke syntetizujú proteíny procesmi, ktoré vedú od genetickej informácie uloženej v molekulách DNA (nukleotidové bázy) k proteínom.

Trojica nukleotidových báz sa nazýva *kodón*. Kodóny sa prekladajú na aminokyseliny, ktoré tvoria druhú úroveň genetického kódu (20 písmenová abeceda). Funkcia proteínu závisí od chemických vlastností aminokyselín, z ktorých sa skladá, ako aj od trojrozmernej štruktúry, ktorú nadobudne poskladaný reťazec aminokyselín v priestore.

1.1 Štruktúra proteínov a proteínové domény

Primárna štruktúra proteínu predstavuje aminokyseliny, z ktorých sa skladá. Pod *sekundárnou štruktúrou* rozumieme popísanie tejto sekvencie z hľadiska lokálnych vodíkových väzieb medzi susediacimi aminokyselinami. Vďaka týmto väzbám nadobúdajú časti sekvencie pravidelný tvar (závitnica α-helix alebo vyrovnaný β-list). Identifikovanie sekundárnej štruktúry proteínu spočíva v označení takýchto úsekov v primárnej štruktúre. Pozície jednotlivých atómov v aminokyselinách proteínu v trojrozmernom priestore určujú *terciárnu štruktúru*.

Doména proteínu je úsek sekvencie aminokyselín, ktorý zodpovedá za určitú chemickú

vlastnosť proteínu alebo funkciu, ktorú proteín vykonáva (napríklad viazať, fyzicky interagovať s inými proteínmi alebo viazať sa na jednovláknovú molekulu DNA). Úseky proteínu, ktoré obsahujú dôležitú doménu, sú menej náchylné na zmenu v priebehu evolúcie, pretože často kódujú kľúčové vlastnosti proteínu.

1.2 Evolúcia, ortológy

Počas evolúcie dochádza mutáciami k postupným zmenám v genetickom kóde, ktoré vedú k vývoju rôznych živočíšnych druhov. Mutácia životne dôležitého génu v organizme by mohla viesť k jeho znefunkčneniu a tým k úmrtiu bunky.

Duplikácia génu je často považovaná za hlavný mechanizmus, ktorý umožnuje diverzifikáciu druhov (Taylor and Raes, 2004). Pri delení bunky sa môže stať, že časť kódujúca niektorý gén sa skopíruje dvakrát a nová bunka bude obsahovať dve kópie tohto génu. Toto umožní kumuláciu mutácií v niektorej z kópií génu pri zachovaní funkčnosti niektorej z kópií. Takýmito zmenami postupne dochádza k vývoju nových živočíšnych druhov z pôvodného spoločného predka. Gény, ktoré vznikli zo spoločného predka, sa nazývajú *homologické*.

Rozlišujeme dva typy homológie medzi génmi, ktoré sa líšia udalosťou, ktorá z pôvodného génu v spoločnom predkovi vytvorila dva rozdielne gény. *Ortológy* sú definované ako gény, ktoré vznikli následkom speciácie dvoch rozdielnych druhov z jedného predka (Fitch, 2000). Na tieto gény sa dá pozerať aj ako na „rovnaké gény“ v rôznych organizmoch a dá sa predpokladať, že vykonávajú rovnakú funkciu. Dva gény sú *paralogické*, ak vznikli z jedného génu jeho duplikáciou (Fitch, 2000)

Zmapovať tieto vťahy medzi génmi, je užitočné, pretože tieto gény si počas evolúcie zachovali niektoré spoločné vlastnosti. To znamená, že informácie, ktoré sú platné o jednom géne z dobre preskúmaného organizmu, sa dajú použiť na predikciu rôznych vlastností jeho ortológu v inom organizme. Spracovať údaje o geneteckých informáciách rôznych organizmoch je kvôli ich objemu náročné, preto je vhodné použiť na ich analýzu informatické metódy.

Kapitola 2

Metódy hľadania ortologických génov

Pretože vzťah ortológie medzi proteínmi vyplýva z ich evolučnej príbuznosti (vyvinuli sa mutáciami z jedného génu v spoločnom predkovi), ortologické proteíny majú často do veľkej miery podobnú genetickú sekvenciu. Pre nájdenie vzťahu ortológie medzi evolučne vzdialenejšimi proteínmi je nutné zobrať do úvahy aj ďalšie dáta ako napríklad pozície génov v ich genóme, či interakcie kódovaných proteínov s ostatnými proteínmi v danom organizme.

Ak sekvencie pochádzajú zo spoločného evolučného predka, dá sa predpokladať, že sa líšia len v určitom počte mutácií, a teda že sú z väčšej časti rovnaké. Počas evolúcie sa v sekvencii predka mohli zmeniť jednotlivé bázy, časti sekvencie mohli byť vypustené alebo naopak mohli byť určité časti do sekvencie pridané.

Pri určovaní podobnosti dvoch sekvencií sa hľadá také ich *zarovnanie*, ktoré medzi nimi určí najväčšiu zhodu. Táto zhoda sa väčšinou vyjadruje cez skóre, ktoré danému zarovnaniu priradí zvolená *skórovacia schéma*. Tieto schémy berú do úvahy dĺžky medzier v sekvencii alebo vlastnosti aminokyselín a s tým spojený dopad zmeny jednej aminokyseliny na inú. Na skórovanie mutácie jednej aminokyseliny na druhú sa používa *skórovacia matica*.

Prvky skórovacej matice popisujú skóre pre zarovnanie každej možnej dvojice písmen z abecedy (pre proteíny 20 aminokyselín) zarovnávaných sekvencií. Riadky aj stĺpce matice predstavujú jednotlivé písmená z používanej abecedy. Skórovacie matice vychádzajú z porovnania pravdepodobnosti výskytu aminokyselín v dvoch evolučne príbuzných sekvenciách a pravdepodobnosti výskytu aminokyselín v náhodnej sekvencii.

Globálne a lokálne zarovnanie

Vo všeobecnosti sa uvažujú dva problémy zarovnania sekvencií. Pri *globálnom zarovnaní* je snaha nájsť podobnosť medzi sekvenciami ako celkami. Zarovnanie len určitých úsekov sa nazýva *lokálne zarovnanie*. Toto zarovnanie sa používa napríklad pri hľadaní jednotlivých génov v sekvencii celého genómu alebo hľadaní domén v proteíne.

Pri hľadaní týchto zarovnaní vychádzame z dobre študovaného informatického problému spoločných podpostupností reťazcov. Rovnako ako pre tento problém sa pri hľadaní oboch typov zarovnania používajú algoritmy založené na princípe dynamického programovania. Needleman-Wunschov a Smith-Watermanov algoritmus (Smith and Waterman, 1981; Needleman and Wunsch, 1970) nájdu optimálne globálne a lokálne zarovnanie pre dané dve sekvencie a dané skórovacie parametre (matica, penalizácia za medzery). Tieto algoritmy sú príliš časovo náročné pre veľký počet dlhších sekvencií.

Sada programov BLAST (Altschul et al., 1990) využíva na hľadanie signifikantných lokálnych zarovnaní podľa danej skórovacej schémy heuristický prístup a snaží sa dosiahnuť optimálne výsledky, ktoré by dosiahli algoritmy dynamického programovania.

Hľadanie ortologických sekvencií

Najpriamočiarejšie použitie BLASTu na hľadanie ortológov je metóda *reciprocal best hit* - najlepších vzájomných hitov (výsledkov hľadania). Podstatou tejto metódy je predpoklad, že dva ortologické gény v dvoch organizmoch, budú navzájom najpodobnejšie, pretože vznikli zo spoločného predka. Gén x z genómu X je najlepším vzájomným hitom (rbh) pre gén y z genómu Y, ak najlepší výsledok hľadania BLASTom s query x v genóme Y je gén y a zároveň pri obrátenom hľadaní s query génu y v genóme X bude najlepší BLAST hit gén x.

Pri analýze BLAST hitov pri RBH metóde je potrebné odlíšiť hľadané ortológy od paralógov, ktoré vznikli duplikáciou ešte pred speciáciou. V genómoch okolo predpokladanej dvojice ortologických proteínov sa očakáva prítomnosť ďalších podobných ortologických dvojíc.Tento predpoklad poskytne ďalšie kritérium pre určenie vzťahu ortológie.

Hľadanie ortologických sekvencií pomocou profilov a motívov

Na určenie úsekov zachovaných medzi viacerými sekvenciami sa používa *viacnásobné zarovnanie sekvencií* (nástroje CLUSTALW (Larkin et al., 2007) a MUSCLE (Edgar, 2004)). Z výstupov týchto nástrojov sa dajú identifikovať zachované úseky medzi jednotlivými sekvenciami. Každý stĺpec zarovnania predstavuje prvky (prípadne medzery) vyskytujúce sa na danom mieste v jednotlivých sekvenciách. Zhoda prvkov v stĺpci pre väčšinu sekvencií naznačuje, že aj pri ďalších sekvenciách patriacich do rovnakej skupiny sa dá očakávať na danej pozície práve rovnaký prvok, ako u väčšiny zarovnaných sekvencií. Na reprezentovanie spoločných čŕt skupiny génov sa následne používajú štruktúry, ktoré umožňujú rozdielne skórovanie zhody medzi danou sekvenciou a modelom v závislosti na pozícii. Na dosiahnutie takejto reprezentácie sa používajú napríklad pozične špecifické skórovacie matice (*position specific scoring matrix - PSSM*) (Gribskov et al., 1987) alebo *profilové skryté Markovovské modely* (*profile Hidden Markov Model - HMM*) (Krogh et al., 1994).

Jedno z možných použití HMM je, že nastavením týchto parametrov sa dajú vytvoriť modely, ktoré reprezentujú určitú skupinu sekvencií. Profilové HMM sú špeciálny prípad skrytých Markovovských modelov, ktoré svojou architektúrou umožňujú z viacnásobného zarovnania vyjadriť spoločné vlastnosti evolučne príbuzných biologických sekvencií. Následne sa takýto model dá použiť na vypočítanie pravdepodobnosti príslušnosti danej sekvencie k modelovanej skupine tak, že sa vypočíta pravdepodobnosť vygenerovania tejto sekvencie týmto modelom. Na základe tejto pravdepodobnosti sa dá posúdiť, či daná sekvencia patrí do skupiny príbuzných sekvencií, ktoré boli použité na tvorbu modelu. Táto pravdepodobnosť sa dá získať štandardným *forward* algoritmom pre HMM (Durbin et al., 1998). Pre posúdenie relevantnosti danej sekvencie je nutné porovnať výslednú pravdepodobnosť s pravdepodobnosťou vygenerovania danej sekvencie náhodným modelom.

Široko používaným nástrojom je napríklad HMMER (Eddy, 1998), ktorý poskytuje celý rad možností ako použiť profilové HMM (hľadať hity pre HMM, emitovať z HMM sekvencie atď). Použítím tohto nástroja bola tiež vytvorená databáza Pfam (Finn et al., 2007), ktorá obsahuje modely rodín príbuzných proteínov, vytvorených z viacnásobných zarovnaní.

Kapitola 3

Hľadanie domén kombináciou sekvenčných a štrukturálnych motívov

Pre zachovanie biologickej funkcie môže byť dôležitejšia trojdimenzionálna štruktúra, do ktorej sa proteín poskladá, než samotné aminokyseliny tvoriace sekvenciu proteínu. V proteíne, ktorý je poskladaný v priestore, interagujú navzájom aj aminokyseliny, ktoré sú od seba značne vzdialené v rámci sekvencie. Takéto väzby môžu spájať jednotlivé úseky v sekundárnej štruktúre do väčších útvarov (napríklad viac β-listov spojených do β-barelu). Takéto útvary v štruktúre proteínu sa nazývajú *štrukturálne motívy* a ich identifikovanie umožňuje zvýšiť citlivosť vyhľadávania domén so silnejšie zachovanou štruktúrou.

Z dostupných informácií môžu odborníci ručne zostaviť deskriptor, reprezentáciu domény, ktorá sa zameria na dôležité vlastnosti štrukturálneho motívu. Deskriptor nebude vhodný pre každú doménu, zvýšenú citlivosť hľadania môže priniesť tento prístup pre domény, v ktorých je štruktúra výraznou charakteristikou a sekvencia nie je dostatočne konzervovaná pre použitie štandardných metód.

Teloméry sú nukleoproteínové komplexy umiestnené na oboch koncoch lineárnych chromozómov. Pri porovnaní s inými proteínovými sekvenciami medzi druhmi je možné pozorovať rýchlu evolúciu týchto proteínov napriek tomu, že vykonávajú v bunkách rovnakú esenciálnu úlohu a očakávateľná by bola vyššia zachovanosť sekvencií (Linger and Price, 2009). Rýchla evolúcia týchto proteínov znižuje šance identifikácie ortológov týchto proteínov v iných organizmoch len na základe sekvenčnej podobnosti. Na týchto skupine týchto

proteínov sme testovali citlivosť metód pracujúcich so sekvenčnou podobnosťou.

Výsledky našej štúdie ukazujú, že metódy založené na sekvenčnej podobnosti sú vhodné na automatické identifikovanie ortológov pre veľké množstvo hľadaných proteínov. V určitých prípadoch však tieto metódy nenájdu žiadne ortológy napriek tomu, že ich prítomnosť je predpokladaná na základe dôležitosti biologickej funkcie hľadaného proteínu. Príkladom takýchto proteínov sú napríklad CDC13 a EST3, ktoré obsahujú OB-fold doménu. Táto doména zaujíma v priestore charakteristickú štruktúru a uvedené výsledky naznačujú vhodnosť domény OB-fold ako relevantného kandidáta pre vytvorenie deskriptora charakterizujúceho štrukturálne vlastnosti hľadanej domény.

Deskriptor pre proteínové domény

Deskriptor pre proteínové domény je vytvorený rozdelením aminokyselinovej sekvencie proteínu na segmenty, ktoré popisujú základné bloky v sekundárnej štruktúre domény. Základom deskriptora teda je postupnosť segmentov, ktoré zodpovedajú úsekom sekvencie s určitou sekundárnou štruktúrou (α-helix, β-list alebo žiadna štruktúra (*coil*). Pre každý z týchto segmentov je potrebné definovať minimálnu a maximálnu dĺžku. V prípade, že segment obsahuje dobre zachovaný sekvenčný motív, môžeme preň špecifikovať PSSM maticu. Druhá časť deskriptoru popisuje hydrogénové väzby, ktoré vznikajú medzi β-listami po poskladaní sa v priestore. Tieto väzby reprezentujeme *hranami* medzi dvoma segmentami typu B, pričom každá takáto hrana má určený smer, mená interagujúcich segmentov a dĺžku. Smer hrany (paralelný alebo antiparalelný) je určený podľa vzájomného smeru dvoch interagujúcich β-listov. Jedna aminokyselina môže vystupovať najviac v jednej väzbe. Medzi dvoma aminokyselinami patriacimi k rovnakej hrane je vždy práve jedna aminokyselina. Takto definovaný deskriptor stanovuje minimálnu dĺžku sekvencie, ktorá obsahuje doménu: súčet minimálnych dĺžok pre všetky segmenty. Okrem tohto obmedzenia nie sú v deskriptore ďalšie striktné podmienky a deskriptor môžeme zarovnať k ľubovoľnej vstupnej sekvencii aminokyselín.

Zarovnaním deskriptora ku sekvencii rozumieme určenie začiatočného bodu pre deskriptor na sekvencii, ďalej určenie rozmeru pre každý segment deskriptora, určenie začiatočného

bodu pre každý sekvenčný motív špecifikovaný pre segmenty a začiatočný pár aminokyselín pre každú hranu.

S daným deskriptorom a skórovacou schémou, ktorá určí ako veľmi vstupná sekvencia vyhovuje očakávaniam vyplývajúcim zo zarovnania deskriptora na sekvenciu, je možné nájsť najlepšie zarovnanie k vstupnej sekvencii a skóre tohto zarovnania použiť na rozhodnutie, či vstupný proteín obsahuje doménu popísanú deskriptorom.

Skórovanie zarovnania deskriptoru ku sekvencii

Deskriptor špecifikuje vlastnosti primárnej, sekundárnej a terciárnej štruktúry popisovanej domény. Skóre zarovnania S je preto určené lineárnou kombináciou týchto troch zložiek.

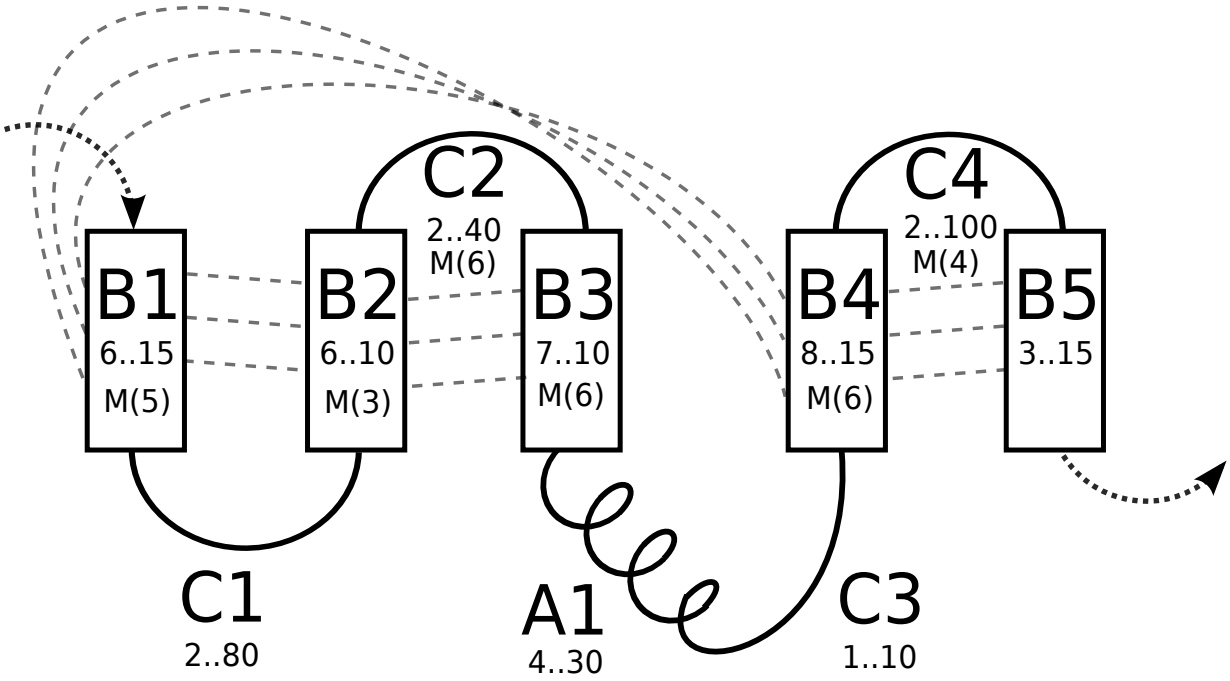
$$S = a * SecStr + b * SeqMot + c * HydBond$$

, kde a, b a c sú váhy pre jednotlivé zložky skóre.

- **Skóre za sekundárnu štruktúru SecStr** - Táto zložka hodnotí, ako dobre korešpondujú typy sekundárnej štruktúry segmentov s predikovanou štruktúrou vstupného proteínu na mieste, kde bol zarovnaný deskriptor.

- **Skóre za sekvenčné motívy SeqMot** - Táto zložka skóre vyjadruje zhodu medzi vstupnou sekvenciou a sekvenčnými motívmi špecifikovanými pre niektoré segmenty. Tieto motívy predstavujú podmienky na očakávanú primárnu štruktúru sekvencie, ktorá by mala obsahovať doménu popísanú deskriptorom.

- **Skóre za hydrogénové väzby HydBond** - Táto zložka skóre hodnotí kompatibilitu segmentov vystupujúcich v hranách deskriptora vytvoriť hydrogénové väzby, ktoré sa po poskladaní vytvoria v doméne popisovanej deskriptorom. Pre skompletovanie skórovacej schémy potrebujme spôsob ako ohodnotiť páry interagujúcich aminokyselín, čo je obsahom kapitoly 4.



Obr. 3.1: Nákres deskriptoru pre OB-fold doménu.

Deskriptor pre Telo_bind doménu

Popísaný typ deskriptoru sme vytvorili pre *Telo_bind* doménu z rodiny OB-foldových domén, do ktorej patria proteíny CDC13 a EST3 z prípadovej štúdie pre telomérické proteíny. Táto doména, podobne ako ostatné z rodiny OB-fold, sa skladá z piatych β-listov sformovaných do β-barelu, ktorý je zväčša uzavretý α-helixom (Arcus, 2002) a β-listy v barele sú zapojené v poradí (1-2-3 5-4-1) (Theobald et al., 2003).

Na obrázku 3.1 je zobrazená schéma deskriptora pre Telo_bind doménu, ktorý má 5 β segmentov, jeden α segment a 4 coil segmenty. Sekvenčné motívy boli na základe HMM modelu pre rodinu Telo_bind z databázy Pfam (Finn et al., 2007) definované pre 6 segmentov (4 β-listy a 2 coil segmenty), ktoré obsahovali úseky s veľkou zhodou medzi doménami v zarovnaní. Väzby a ich dĺžky boli odvodené na základe analýzy 3D modelu Telo_bind domény v proteíne CDC13 z kvasinky *Saccharomyces Cerevisiae* uloženého v databáze PDB (Rose et al., 2011). Väzby medzi jednotlivými segmentami sú v schéme zobrazené šedými prerušovanými čiarami.

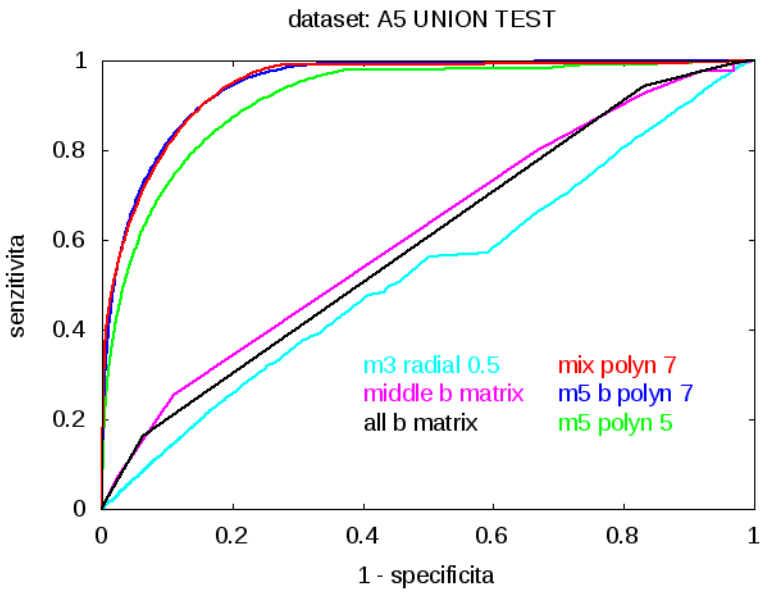
Kapitola 4

Rozpoznávanie aminokyselín spojených hydrogénovou väzbou pomocou SVM

Pre účely našej deskriptorovej metódy sme vytvorili klasifikátor, ktorý ohodnotí ľubovoľnú dvojicu aminokyselín v proteíne z hľadiska očakávanej tendencie vytvoriť väzbu. Pre naše analýzy sme použili údaje o štruktúre 101642 proteínov z databázy Protein Data Bank (Rose et al., 2011). V sekvenciách sme hľadali trojice spárovaných aminokyselín v β-listoch, ktoré nasledovali v sekvencii striedavo za sebou. Takéto trojice párov sa dajú reprezentovať párom úsekov sekvencie dĺžky 5, kde vo väzbách vystupujú prvá, tretia a piata aminokyselina v úseku. Na základe týchto údajov sme natrénovali dve skórovacie matice, ktoré vyjadrujú pravdepodobnosť, že dve aminokyseliny budú interagovať. Každý z SVM modelov bol natrénovaný pre paralelný a antiparalelný smer, s polynomiálnym a radiálnym kernelom s rôznymi parametrami.

Podľa trénovacích sád a použitých atribútov môžeme rozdeliť vytvorené SVM klasifikátory na tri skupiny:

- **SVM pre spárované trojice aminokyselín (M5)** - Príklady v trénovacích množinách sú páry úsekov sekvencie s dĺžkou päť aminokysein. V pozitívnych príkladoch obsahovali tri hydrogénové väzby medzi aminokyselinami v týchto dvoch úsekoch. V tomto prípade bol vytvorený aj model označený *M5 MIX*, ktorý obsahuje paralelné aj antiparalelné prípady.
- **SVM pre spárované trojice aminokyselín s vynechanými pozíciami (M3)** -



Obr. 4.1: Porovnanie ROC kriviek pre natrénované SVM modely a skórovacie matice.

Vynechali sme zo vstupov nespárované aminokyseliny a vstupy v týchto modeloch kódujú iba tri interagujúce páry aminokyselín.

- **SVM pre spárované aminokyseliny so spárovaným susedom (M5B)** - Tieto modely obsahujú medzi pozitívnymi príkladmi aj okrajové príklady, to znamená páry sekvencií, v ktorých sú len dve väzby.

Najlepšie klasifikátory z jednotlivých skupín sme ďalej porovnali pomocou ROC krivkiek. Výsledky tohto porovnania sú zobrazené na obrázku 4.1. Z porovnania vyplýva, že obe skórovacie matice, rovnako ako klasifikátor zložený z M3 modelov, nemajú dostatočnú klasifikačnú silu na odlíšenie pozitívnych príkladov od negatívnych. Výsledky klasifikátora zloženého z modelov M5 sú výrazne lepšie. Klasifikátory s modelmi M5 B, ktoré boli natrénované na sadách rozšírených o okrajové prípady, majú ešte mierne lepšie výsledky. Podobné výsledky dosiahol aj model M5 MIX. Na základe tohto porovnania sme sa rozhodli použiť pre skórovanie očakávaných hydrogénových väzieb medzi aminokyselinami SVM modely M5 B s polynomiálnym kernelom s parametrom delta = 7.

Kapitola 5

Zarovnávanie deskriptorov pomocou celočíselného lineárneho programovania

Pre nájdenie domény pomocou deskriptora musíme nájsť jeho najlepšie zarovnanie k vstupnej sekvencii poďla danej skórovacej schémy. Jedná sa teda o optimalizačnú úlohu, v ktorej maximalizujeme skóre zarovnania a dá sa riešiť jednou zo štandardných metód: celočíselným lineárnym programovaním.

Úlohou je určiť, ktoré aminokyseliny sekvencie zodpovedajú štrukturálnym segmentom a sekvenčným motívom v deskriptore, ako aj pozíciu hrán rprezentujúcich interakcie medzi segmentami. Takéto zarovnanie deskriptora k vstupnej sekvencii popíšeme štyrmi sadami binárnych premenných:

- **Premenné pre umiestnenie segmentov** x_{ij}: premenná x_{ij} má hodnotou 1 práve vtedy, keď na pozícii i v sekvencii je zarovnaný j-ty segment deskriptora.
- **Premenné pre umiestnenie sekvenčných motívov** m_{ij}: premenná m_{ij} má hodnotu 1 práve vtedy keď, na pozícii i v sekvencii začína sekvenčný motív j-teho segmentu deskriptora.
- **Premenné pre spárované segmenty** y_{ijkl}, z_{ijkl}: premenná y_{ijkl} má hodnotu 1 práve vtedy, keď medzi segmentami j a l je paralelná hrana, ktorá začína párom aminokyselín i (v j-tom segmente) a k (v l-tom segmente).
- **Pomocné premenné pre kontrolu prekrývania väzieb** p_{ijl}: Podmienky s týmito

premennými zaručia, že jedna aminokyselina nebude vystupovať v dvoch rôznych väzbách.

Účelová funkcia binárneho programu vygenerovaného pre daný deskriptor a vstupnú sekvenciu má tvar:

$$\sum_{i,j}(e_{ij}x_{ij} + f_{ij}m_{ij}) + \sum_{i,j,k,l}(g_{ijkl}y_{ijkl} + h_{ijkl}z_{ijkl})$$

,

kde e_{ij} je skóre pre zhodu medzi očakávanou sekundárnou štruktúrou aminokyseliny i a sekundárnou štruktúrou segmentu j, m_{ij} je skóre zhody sekvenčného motívu segmentu j a sekvencie na úseku začínajúceho aminokyselinou i, g_{ijkl} (h_{ijkl}) je skóre paralelnej (antiparalelnej) hrany medzi segmentami j a l, ktorá začína spárovanými aminokyselinami i a k. Skóre pre páry interagujúcich aminokyselín sa získa z natrénovaných SVM.

Aby uvedené sady binárnych premenných kódovali prípustné zarovnanie deskriptora ku sekvencii, musia premenné zároveň spĺňať sadu podmienok, ktoré zabezpečia, že segmenty budú súvislé a v správnom poradí, budú mať správnu dĺžku a jedna aminokyselina bude patriť do jedného segmentu, každý motív a väzba je správne umiestnená. Tieto podmienky zabezpečujú, že priradenie hodnôt 0 alebo 1 ku sadám premenných zodpovedá zarovnaniu deskriptora k vstupnej sekvencii. Maximalizovaním účelovej funkcie programu dostaneme optimálne zarovnanie vzhľadom na našu skórovaciu schému.

Výsledky

Takýto celočíselný program sme použili nariešenie problému zarovnania deskriptora OB-fold domény ku sade proteínových sekvencií, ktorá obsahovala 10 proteínov s OB-fold doménou a 13 náhodne vybraných proteínov bez tejto domény. Pre účely testovania celočíselného programu sme z proteínov vybrali úseky sekvencie obsahujúce OB-fold doménu s dĺžkou menšou než 200 (pre negatívne proteíny boli použité náhodné úseky danej dĺžky).

Pre vygenerovanie celočíselného programu pre daný deskriptor a vstupnú sekvenciu sme vytvorili program v jazyku *Java*, ktorý z deskriptora pre OB-fold domény, sekvencie

proteín	dĺžka	P/N	skóre	čas
domain_A2QSY5	155	N	12.863	3d 14h 42m 39s
domain_Q9MUM1	155	N	12.2327	18h 20m 9s
domain_C6DDH0	155	N	-1.0506*	>18d
domain_P23180	155	N	6.2340	16d 21h 29s
domain_B0UU36	155	N	20.1410	6h 55m 27s
domain_B7LAR7	155	N	6.7946	1d 14h 6m 59s
domain_Q230X8	155	N	2.6504	3d 2h 57m 24s
domain_Q2YMQ2	155	N	-1.8042	1d 12h 34m 9s
domain_Q5M1N8	120	N	3.5389	1d 8h 6m 43s
domain_Q72U22	155	N	-0.6774*	>18d
domain_B0K889	155	N	21.7818	15d 1h 34m 41s
domain_Q90229	155	N	8.3781	18d 22h 17m 3s
domain_Q5A455	155	N	29.9385	1h 21s
domain_TEBH_EUPCR	155	P	44.3315	32m 55s
domain_POT1_SCHPO	146	P	39.355	36m 41s
domain_POTE1_CHICK	131	P	44.7688	4m 18s
domain_POTE1_HUMAN	131	P	45.4261	3m 12s
domain_POTE1_MOUSE	131	P	43.6156	5m 42s
domain_TEB EUPCR	120	P	47.3266	47s
domain_TEBA_OXYNO	160	P	51.1779	22m 43s
domain_TEBA_STYMY	160	P	51.2706	29m 47s
domain_CDC13_YEAST	195	P	54.7663	2m 12s
domain_POTE1_MACFA	131	P	45.2819	2m 47s

Tabuľka 5.1: Výsledky pre hľadanie zarovnania pomocou ILP. Stĺpec P/N vyjadruje, či bola skúmaná doména pozitívnym príkladom alebo nie. V niektorých prípadoch sa výpočet neskončil za väčší počet dní. Znak * pri skóre indikuje najlepšie nájdené celočíselné riešenie za daný čas.

aminokyselín s predikovanými pravdepodobnostami pre jednotlivé typy sekundárnej štruktúry a natrénovaných SVM modelov vygeneroval podmienky k celočiselnému programu pre zarovnanie a vypočítal koeficienty v účelovej funkcii. Následne sme takto vygenerovaný celočíselný program riešili solverom CPLEX.

Výsledky pre najlepšie zarovnanie sú v tabuľke 5.1 a rozdiel medzi skóre, ktoré dosiahli pozitívne a negatívne príklady je dostatočne veľký na korektnú separáciu pozitívnych a negatívnych príkladov. Avšak časová náročnosť výpočtu je veľmi veľká, obzvlášť pri negatívnych príkladoch.

Keďže na tento problém narazíme už pri testoch na krátkych úsekoch proteínov, dlhý čas výpočtu znemožňuje praktické použitie celočíselného programovania na hľadanie OB-fold domény v celých proteínoch. Z tohto dôvodu sme museli nájsť výrazne rýchlejšiu metódu na určovanie najlepšieho zarovnania.

Kapitola 6

Zarovnávanie deskriptorov pomocou dynamického programovania

6.1 Relaxovaný problém

Všeobecne formulovaný problém zarovnania deskriptoru ku sekvencii môže popisovať ľubovoľné interakcie medzi vzdialenými segmentami, pričom pozície spárovaných aminokyselín nemajú žiadne ďalšie obmedzenia. Riešenie takto všeobecne formulovaného problému teda môže byť časovo veľmi náročné. Ak zavedieme ďalšie obmedzenia na konfiguráciu interakcií medzi segmentami popísanými v deskriptore, je možné využiť toto zúženie problému a nájsť rýchlejšie riešenie pre určitú množinu deskriptorov.

Ak medzi dvoma segmentami s_1 a s_2 existuje v deskriptore hydrogénová väzba, budeme túto skutočnosť označovať $s_1 \sim s_2$. Ak segment s_1 leží v deskriptore s_2, označujeme túto skutočnosť ako $s_1 < s_2$. Segment s je *osamelý*, ak nevystupuje v žiadnej hydrogénovej väzbe. Deskriptor spĺňa *obmedzenie na konfiguráciu väzieb* ak pre každé dva segmenty $s_1 < s_2$, pre ktoré $s_1 \sim s_2$, platí, že všetky segmenty s ležiace medzi s_1 a s_2 ($s_1 < s < s_2$) sú osamelé.

Deskriptory, ktoré uvažujeme v relaxovanom probléme zarovnania, spĺňajú obmedzenie a teda vyjadrujú len jednoduchšiu štruktúru väzieb medzi segmentami. Tieto štruktúry tvoria menšie navzájom nezávislé oblasti deskriptora, čo umožňuje riešiť problém zarovnania deskriptora pomocou dynamického programovania. Deskriptor pre OB-fold domény popísaný v kapitole 3 nespĺňa obmedzenie , pretože väzba $B1 \sim B4$ pokrýva väzbu

$B2 \sim B3$ a má spoločný začiatok s väzbou $B1 \sim B2$. Z tohto dôvodu budeme v dynamickom programovaní používať deskriptor, ktorý neobsahuje túto väzbu. Týmto spôsobom dostaneme odhad optimálneho riešenia pre kompletný deskriptor.

Dynamické programovanie

Najlepšie zarovnanie zjednodušeného deskriptoru k danej sekvencii dokážeme vypočítať v polynomiálnom čase prostredníctvom dynamického programovania. Dynamickým programovaním budeme riešiť problém, ako dostať najlepšie zarovnanie pre prvých n segmentov deskriptora, za predpokladu, že poznáme najlepšie zarovnanie pre prvých $n-1$ segmentov. Matica dynamického programovania bude mať tvar $A[n, k, f', k', r]$ a bude obsahovať skóre pre najlepšie zarovnanie prvých n segmentov končiace na pozícii k. Parametre f', k' a r sú potrebné pre počítanie skóre pre hydrogénové väzby, ktoré spájajú vzdialené aminokyseliny.

Dynamické programovanie pre problém zarovnania deskriptoru ku sekvencii pre zjednodušené deskriptory má časovú zložitosť $O(nmfl^4)$, kde n je dĺžka vstupnej sekvencie, m je počet segmentov, f je dĺžka maximálneho rozdielu medzi dvoma interagujúcimi segmentami a l je maximálna dĺžka segmentu v deskriptore. Pamäťová zložitosť tohto algoritmu je $O(nmfl^2)$ je určená rozmermi matice dynamického programovania.

Orezávanie

Skórovanie potenciálnych párov interagujúcich aminokyselín pomocou SVM je časovo najnáročnejšou časťou nášho algoritmu. Aminokyseliny, ktoré nepatria do žiadneho β-listu, pravdepodobne nebudú súčasťou niektorej z väzieb medzi β-listami. Ak nebudeme vyhodnocovať páry, kde je aspoň jedna takáto aminokyselina, výrazne obmedzíme počet vstupov do SVM klasifikátora, a teda aj čas potrebný na výpočet zarovnania.Preto sme zaviedli orezávacie pravidlo, ktoré vyradí z vyhodnocovania SVM páry aminokyselín, kde aspoň jedna aminokyselina má predikovanú pravdepodobnosť príslušnosti do β-listu menšiu ako 50 %.

proteín	dĺžka	P/N	skóre bez orezávania	skóre s orezávaním
A2QSY5	281	N	18.8356	18.4966
Q9MUM1	446	N	13.3688	orezané
C6DDH0	302	N	17.5898	orezané
P23180	465	N	34.2771	orezané
B0UU36	334	N	20.6194	orezané
B7LAR7	200	N	15.5853	orezané
Q230X8	423	N	27.5368	27.1412
Q2YMQ2	435	N	5.4542	orezané
Q5M1N8	241	N	13.5218	orezané
Q72U22	307	N	7.2892	orezané
B0K889	233	N	21.3008	21.27
Q90229	280	N	61.1585	orezané
Q5A455	762	N	44.3158	44.3154
TEBH_EUPCR	460	P	45.9458	45.9458
POT1_SCHPO	555	P	44.3967	44.3967
POTE1_CHICK	778	P	55.247	55.247
POTE1_HUMAN	634	P	45.5979	45.5979
POTE1_MOUSE	640	P	42.3972	42.3972
TEB EUPCR	420	P	49.2403	49.2403
TEBA_OXYNO	495	P	52.2969	52.2398
TEBA_STYMY	493	P	52.5483	52.3672
CDC13_YEAST	924	P	58.9341	58.7209
POTE1_MACFA	634	P	46.7132	46.7132

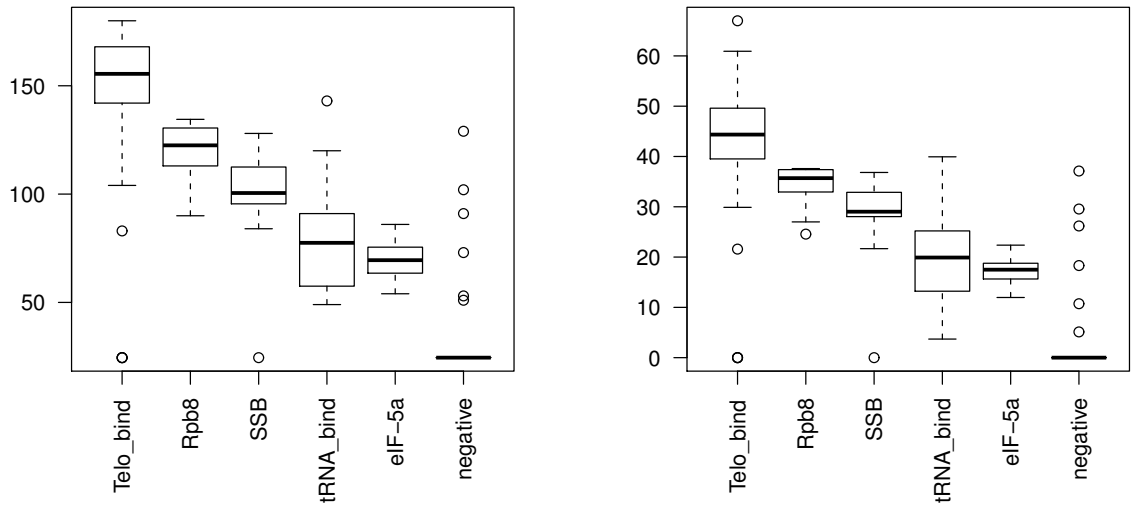
Tabuľka 6.1: Tabuľka popisuje skóre najlepšieho zarovnania (s orezávaním a bez neho) deskriptoru pre Telo_bind doménu (viď kapitola 3) ku sekvenciám 23 proteínov. Stĺpec P/N indikuje,či je daný proteín pozitívnym príkladom (P) a je v ňom prítomná doména popísaná deskriptorom alebo či daný proteín takúto doménu nemá (N).

Experimentálne overenie metódy

Klasifikačnú silu hľadania špecifických domén v proteínoch pomocou deskriptora sme overili na reálnych dátach. Dynamickým programovaním sme hľadali doménu Telo_bind z rodiny domén OB-fold. Dynamické programovanie umožňuje v kratšom čase spočítať zarovnanie aj pre dlhšie sekvencie, zarovnávali sme deskriptor k celým proteínom a nie len k vybraným úsekom proteínu.

Na skóre zarovnaní opäť vidno dobrú separáciu negatívnych a pozitívnych príkladov výskytu hľadanej domény v proteíne. Skóre zarovnania s orezávaním pozícií, ktoré majú malú pravdepodobnosť príslušnosti k β-listu môžeme porovnať so skóre zarovnania bez tohto orezávania. Vidíme, že rozdiely medzi týmito skóre sú minimálne a v niektorých prípadoch obe metódy nájdu to isté zarovnanie.

Pri ďalšom teste sme z databázy Pfam (Finn et al., 2007) vybrali 50 proteínov, ktoré obsahujú hľadanú doménu Telo_bind a 50 proteínov bez nej. Následne sme z rovnakej



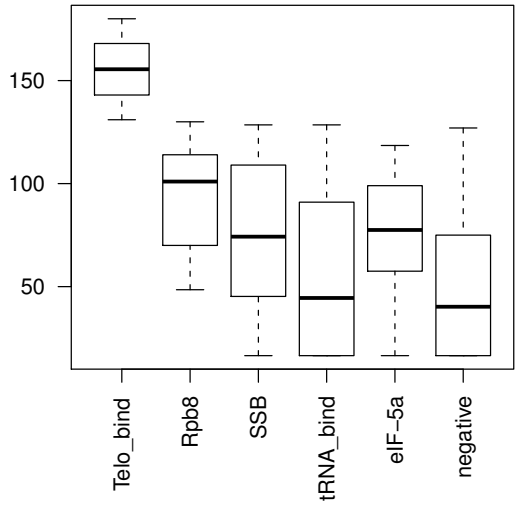
Obr. 6.1: Výsledky pre hľadanie Telo_bind deskriptor. Na osi x sú jednotlivé skupiny proteínov použité v testovacej množine. Na grafe vľavo sú na osi y je umiestnenie proteínov (rank) z danej skupiny vo výsledkoch hľadania. Na grafe v pravo je na osi y vyznačené skóre, ktoré dosiahli zarovnanie deskriptoru na jednotlivých proteínoch z danej skupiny.

databázy vybrali 20 proteínov pre štyri rodiny príbuzné k Telo_bind. Takýmto spôsobom sme zostavili množinu 180 proteínov rôznej dĺžky, na ktorých sme deskriptorom hľadali Telo_bind doménu.

Výsledky tohto hľadania sú zobrazené na obrázku 6.1. Z výsledkov vyplýva, že táto metóda dokáže spoľahlivo separovať negatívne príklady od proteínov obsahujúcich OB-fold domény a zároveň dokáže odlíšiť špecifickú doménu Telo_bind od príbuzných rodín s OB-fold doménami. Proteíny s príbuznou doménou, ktorá nepatrila do rodiny Telo_bind majú skóre blízko k Telo_bind proteínom, čo je výhoda, pretože to indikuje možnosť využitia deskriptora na hľadanie vzdialených ortológov a proteínov s príbuznou doménou.

6.1.1 Porovnanie s programom Hmmer

Klasifikačnú silu našej metódy sme porovnali s nástrojom Hmmer (Eddy, 2011), pomocou ktorého bola vytvorená databáza Pfam. V prípade OB-fold proteínov z rodín vzdialenejšie príbuzných k rodine Telo_bind nástroj Hmmer nedetekuje podobnosť a vo výsledku sú



Obr. 6.2: Výsledky pre hľadanie OB Fold domén pomocou Hmmeru. Na osi x sú jednotlivé skupiny proteínov použité v testovacej množine. Na osi y je umiestnenie proteínov (rank) z danej skupiny vo výsledkoch hľadania.

tieto proteíny nerozlíšiteľné od negatívnych príkladov proteínov bez OB-fold domény. Toto poukazuje na výhodu nášho prístupu, ktorý ich dokáže odlíšiť od negatívnych príkladov.

6.1.2 Celogenómové hľadanie OB-Fold proteínov

Ako ďalší test sme skúsili hľadať proteíny s Telo_bind doménov v proteóme kvasinky *Yarrowia lipolytica*. Deskriptor pre OB-fold doménu sme zarovnali k 6292 proteínom z kvasinky *Yarrowia lipolytica*. Veľká väčšina (4656) z týchto proteínov bola orezaných vyššie popísanou podmienkou ako proteíny, na ktoré sa nedá deskriptor zarovnať. Pre zvyšných 1636 proteínov bolo vypočítané najlepšie zarovnanie.

Výsledky sme usporiadali podľa zložky skóre za hydrogénové väzby a sústredili sme sa na zarovnania, ktoré sú v prvej stovke najlepšie hodnotených proteínov podľa tejto zložky, a zároveň patria do množiny 80 proteínov s celkovým skóre viac ako 38. Týmto spôsobom sme dostali 14 proteínov vhodných na ďalšiu analýzu. Výsledky experimentu demonštrujú, že metóda hľadania štrukturálnych domén pomocou deskriptora je aplikovateľná na úrovni celého genómu a môže slúžiť ako filter pre veľké množstvo sekvencií.

Záver

Na riešenie problému identifikovania ortologických proteínov sme definovali reprezentáciu domén pomocou ručne zostavených deskriptorov, ktoré kombinujú viaceré typy informácií o štruktúre popisovanej domény. Na riešenie problému zarovnania sme najskôr navrhli celočíselný lineárny program. Zarovnania proteínov, ktoré obsahovali doménu popísanú deskriptorom, dosahovali výrazne vyššie skóre, ako proteíny bez hľadanej domény, takže táto metóda dokázala identifikovať prítomnosť hľadanej domény v proteínoch. Avšak ohodnotenie negatívnych príkladov bolo príliš časovo náročné, výpočet pre veľkú časť z týchto príkladov trval niekoľko dní. Použitie tejto metódy je teda nepraktické v prípadoch, kedy je potrebné zarovnať deskriptor k veľkému množstvu sekvencií.

Problém zarovnania sme zjednodušili zavedením obmedzenia na prípustné väzby v deskriptore. Tento zjednodušený problém sme riešili pomocou dynamického programovania, ktoré dokázalo nájsť zarovnanie deskriptora aj k dlhším sekvenciám v prijateľnom čase. Vďaka tomuto zjednodušeniu je možné aplikovať hľadanie deskriptorom aj na úrovni celého genómu.

Pomocou tohto algoritmu sme otestovali deskritor pre Telo_bind doménu na sade 180 proteínov a porovnali sme naše výsledky s programom Hmmer, ktorý používa na reprezentáciu domény profilové HMM. Metóda hľadania domén pomocou deskriptora dokáže korektne rozlíšiť proteíny s doménou Telo_bind od proteínov bez tejto domény. Navyše výhodou našej metódy je jej schopnosť identifikovať aj proteíny s doménami, ktoré sú príbuzné k hľadanej doméne Telo_bind. Túto metódu je možné použiť na analýzu celých genómov, čo sme demonštrovali na genóme kvasinky *Yarrowia lipolytica*.

Literatúra

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Arcus, V. (2002). Ob-fold domains: a snapshot of the evolution of sequence, structure and function. *Current opinion in structural biology*, 12(6):794–801.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis*. Cambridge Univ. Press.

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9):755.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol*, 7(10):e1002195.

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792.

Finn, R., Tate, J., Mistry, J., Coggill, P., Sammut, S., Hotz, H., Ceric, G., Forslund, K., Eddy, S., Sonnhammer, E., et al. (2007). The Pfam protein families database. *Nucleic acids research*.

Fitch, W. (2000). Homology:: a personal view on some of the problems. *Trends in genetics*, 16(5):227–231.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 84:4355–4358.

Krogh, A., Brown, M., Mian, I., Sjolander, K., and Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–1531.

Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947.

Linger, B. and Price, C. (2009). Conservation of telomere protein complexes: shuffling through evolution. *Critical Reviews in Biochemistry and Molecular Biology*, (00):1–13.

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B., Zardecki, C., Berman, H. M., and Bourne, P. E. (2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res*, 39(Database issue):D392–401.

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.

Taylor, J. and Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas.

Theobald, D., Mitton-Fry, R., and Wuttke, D. (2003). Nucleic acid recognition by ob-fold proteins. *Annual review of biophysics and biomolecular structure*, 32:115.

UNIVERZITA KOMENSKÉHO

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Zoznam publikačnej činnosti

Mgr. Martin Macko

AFK Postery zo zahraničných konferencií

AFK01 M. Macko [UKOMFKAI] - Ľ. Tomáška - T. Vinař [UKOMFKAI]: Orthologous proteins associated with yeast telomeric complex identified by synteny and sequence similarity
In: ECCB 10 (CD ROM). - Leuven : Katholieke Universiteit, 2010. - S. 21
[ECCB 2010 : European Conference on Computational Biology. 9th, Ghent, 26.-29.9.2010]

Štatistika kategórií (Záznamov spolu: 1):
AFK Postery zo zahraničných konferencií (1)

27.7.2012