SEMINAR
ZA FIZIKO MEHKE SNOVI
Napovedovanje sekundarne strukture beljakovin

Borut Tone Oblak

10. maj 2006

Povzetek

Čeprav obstaja mnogo beljakovin brez stalne tridimenzionalne strukture, je pri večini beljakovin tridimenzionalna struktura vendarle odločilna za njihovo funkcijo. Odkrivanje tridimenzionalnih struktur beljakovin poteka precej počasneje kot odkrivanje njihovih aminokisliskih zaporedij, zato je že dolgo aktualno vprašanje, ali se da na podlagi aminokislinskega zaporedja beljakovine vsaj približno določiti njeno tridimenzionalno strukturo. Algoritmi za napovedovanje celotne tridimenzionalne strukture so šele na začetku svojega razvoja, drugače pa je z algoritmi, ki napovedo le sekundarno strukturo. Ti so danes že dokaj natančni, saj sega njihova zanesljivost čez 70%, kar ni tako malo, če upoštevamo, da je največja možna zanesljivost določitve sekundarne strukture beljakovine 88.4%.

1 Uvod

Tradicionalen pogled na beljakovine predpostavlja, da tridimenzionalna struktura beljakovine pretežno ali v celoti določa biološko funkcijo beljakovine. To ni čisto res, vsaj dobesedno ne, saj približno $10-20\%$ znanih beljakovin nima stalne tridimenzionalne strukture [1], ampak se jim ta ves čas spreminja, kakor se na primer vse čas spreminja oblika niti v vodnem toku. Take beljakovine tridimenzionalne strukture dejansko nimajo in se jim z nobeno metodo ne da določiti niti sekundarne niti terciarne in kvartarne strukture. Ta lastnost je še bolj izrazita, če ne gledamo celotnih beljakovinskih molekul, ampak njihove dele – ocenjuje se, da kar $25-40\%$ do sedaj znane primarne

strukture (aminokislinskih zaporedij) predstavlja področja brez stalne tridimenzionalne strukture. Vendar pa zveza med tridimenzionalno strukturo in funkcijo vendarle ostaja, čeprav prek odsotnosti prve – beljakovine brez stalnih tridimenzionalnih struktur opravljajo tiste funkcije v organizmu, ki od njih zahtevajo stalno spreminjanje oblike. Tak primer je na primer *titin*, beljakovina z velikimi deli brez stalne tridimenzionalne strukture, ki služi kot entropična vzmet v mišičnih celicah.

Kljub povedanemu, pa je informacija o tridimenzionalni strukturi tiste večine beljakovin in njihovih delov, ki stalno 3D strukturo imajo, ključnega pomena za razumevanje njihove funkcije. Žal pa je določanje 3D struktur beljakovin z rentgenskim sipanjem ali NMR zapleteno in počasno, tako da je število znanih 3D struktur beljakovin desetkrat manjše od števila znanih primarnih struktur. Zato se vse bolj uveljavlja računalniško napovedovanje 3D struktur beljakovin. Idealno bi bilo razvozlati procese zvijanja na podlagi prvih principov, ter na podlagi sil, ki nastopajo med deli beljakovine ter beljakovine in topila, določiti tako potek zvijanja kot tudi njegov rezultat – tridimenzionalno strukturo. Vendar je to zelo zahteven problem, tako da je ta pristop bolj uporaben za sam študij procesa zvijanja, kot za napovedovanje tridimenzionalne strukture. Zato so, kot že večkrat v fiziki, ko iskanega niso zmogli zanesljivo izračunati iz prvih principov, uporabili fenomenološki pristop. Na trenutni razvojni stopnji, lahko algoritmi za napovedovanje precej dobro napovedo sekundarno strukturo, napovedovanje terciarne in kvartarne strukture pa je še zelo nezanesljivo. Zato se bom v tem seminarju posvetil algoritmom za določanje sekundarne strukture.

2 Prva in druga generacija algoritmov za napovedovanje sekundarne strukture

Pri napovedovanju sekundarne strukture želimo za vsako aminokislino v aminokislinskem zaporedju določiti v kateri element sekundarne strukture spada (*sekundarno strukturno stanje aminokisline*) – torej, ali je del vijačnice α, lista β, obrata ali pa naključnega navitja.

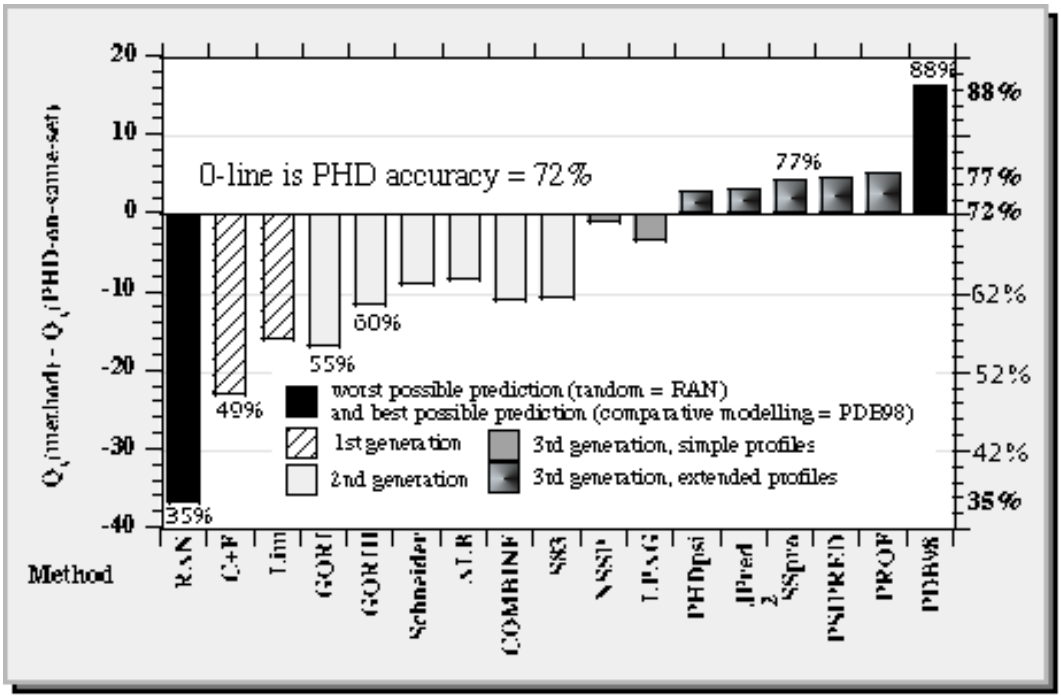
Kmalu po eksperimentalni določitvi prvih tridimenzionalnih (in s tem tudi sekundarnih) struktur so opazili, da so nekatere aminokisline pogostejše v nekaterih elementih sekundarne strukture, druge pa v drugih. Leta 1974 sta Chou in Fasman izračunala tabelo verjetnosti za to, da se aminokislina neke vrste nahaja v nekem elementu sekundarne strukture [2, 3, 4, 5]. S pomočjo te tabele sta potem napovedovala sekundarno strukturo beljakovin. Zanesljivost takšne metode je bila komaj 50% (slika 1). Zanesljivost metode

je razmerje med številom aminokislin, ki jim je metoda pravilno napovedala sekundarno strukturno stanje in številom vseh aminokislin, pri čemer se navadno upoštevajo le tri možna stanja: vijačnica α, list β in vse ostalo. Na podlagi te metode se jih je razvilo še nekaj, vse pa predpostavljajo, da vrsta aminokisline na nekem mestu sama določa sekundarno stukturno stanje te aminokisline. To so *algoritmi 1. generacije*. Nobeden od njih nima zanesljivosti dosti boljše od 50%.

Do pomembnejšega premika je prišlo leta 1974, ko so Greiner, Osguthorpe in Robson [6] razvili algoritem *GOR* po katerem je sekundarno strukturno stanje aminokisline določeno z njeno okolico v aminokislinskem zaporedju. Vzeli so okno 17 zaporednih aminokislin (centralno, 8 levo in 8 desno), ki je drselo po aminokislinskem zaporedju, in statistično analizirali pogostnost pojavljanja vrst aminokislin v njem. Vektor frekvenc pojavljanja aminokislin v oknu so množili z matriko vnaprej fenomenološko določenih uteži. Iz dobljene vrednosti so določili sekundarno strukturno stanje centralne aminokisline. To je že *algoritem 2. generacije*, saj predpostavlja, da je sekundarno strukturno stanje neke aminokisline posledica njene okolice in ne (le) nje same. Zgoraj opisani prvotni algoritem GOR sploh ne prizna nikakršnega neposrednega vpliva centralne aminokisline na svoje sekundarno strukturno stanje, kar pa se je izkazalo kot slabost. Zato novejši algoritem GOR III, ki so ga razvili Gibrat, Greiner in Robson leta 1987 [7] upošteva korelacijo med centralno aminokislino in preostalimi v oknu tako, da namesto ene matrike uteži uporablja 20 matrik, za vsako vrsto centralne aminokisline drugo. Njihove vrednosti so ponovno določene fenomenološko. Zadnji algoritem te skupine GOR IV iz leta 1996 [8], pa upošteva medsebojne korelacije vseh aminokislin v oknu. Zanesljivost GOR III je okoli 60% (slika 1), zanesljivost GOR IV pa malo nad 60%.

Algoritem GOR je le eden od predstavnikov 2. generacije algoritmov. Za vse algoritme te generacije je značilno, da analizirajo segmente aminokislinskega zaporedja beljakovine, dolge navadno od 11 do 21 zaporednih aminokislin. Na podlagi analize pojavnosti aminokislin v segmentu izračunajo verjetnost za to, da je centralna aminokislina v segmentu v nekem sekundarnem strukturnem stanju. In iz teh verjetnosti potem določijo sekundarno strukturo celotne beljakovine. Te verjetnosti različne metode lahko računajo sila različno: ene temeljijo na statistični analizi pojavnosti različnih vrst aminokislin v različnih elementih sekundarne strukture, druge na fizikalno-kemijskih lastnostih aminokislin, pogoste so tudi nevronske mreže, ki jih učijo na že znanih podatkih. Skupno vsem tem algoritmom je, da je njihova zanesljivost največ malce nad 60%.

Poleg sorazmerno nizke zanesljivosti pri algoritmih 1. in 2. generacije obstajata še dve pomembni težavi: še posebej slaba je zanesljivost napovedi



Slika 1: Primerjava zanesljivosti različnih programov za napovedovanje sekundarne strukture vseh treh generacij. Leva navpična skala je relativna zanesljivost glede na program PHD iz 3. generacije, ki ima zanesljivost 72%, desna navpična skala pa je absolutna zanesljivost programa. Stolpec skrajno levo pomeni delež pravilno napovedanih sekundarnih strukturnih stanj, če bi bilo določanje naključno (35.2%, najmanjša možna zanesljivost), stolpec skrajno desno pa največjo možno zanesljivost (88.4%).

listov β, le $28-48\%$, kar je komaj malo boljše, kot če bi jih določili naključno, ali pa še to ne; poleg tega pa so vijačnice α in listi β prekratki.

Razmeroma nizka zanesljivost izhaja iz dveh virov:

1. Ista beljakovina lahko kristalizira na več načinov, ki imajo lahko malce različne elemente sekundarne strukture zato je največja možna zanesljivost napovedi 88.4%. Najmanjša možna zanesljivost je verjetnost za naključno pravilno določitev sekundarnega strukturnega stanja aminokisline, ki je 35.2%.

2. Na sekundarno strukturo vplivajo tudi interakcije med oddaljenimi aminokislinami, torej na sekundarno strukturno stanje centralne aminokisline vplivajo tudi aminokisline, ki jih nismo zajeli v okno.

Nizka zanesljivost napovedovanje listov β, je bila razložena z dejstvom, da na njihovo oblikovanje interakcije med oddaljenimi aminokislinami (v aminokislinskem zaporedju) vplivajo bistveno bolj kot na vijačnice α in ostala elementa sekundarne strukture.

Vzroki za prekratke vijačnice α in listi β so zaenkrat neznani, pomenijo pa recejšnjo težavo pri praktični uporabi rezultatov (slika 2).

Bilo je mnogo poskusov, da bi te težave vsaj resno omilili, če ne že odpravili, a očitno z algoritmi te vrste to ni mogoče, za bistveno povečanje zanesljivosti je bilo potrebno k problemu že v osnovi pristopiti drugače.

```
SEQ   KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVP
OBS     EEEE          E--E    EEEEEE     EEEEEE  EEEEEE
TYP     EEHHH         EE    EEEE   EE    HHHEE       EEE
```

Slika 2: Tipični primer določitve sekundarne strukture z algoritmom 2. generacije (zanesljivost 60%). SEQ je sekvenca - zaporedje aminokislin, ki jim določamo sekundarno strukturo. OBS je eksperimentalno določena sekundarna struktura, TYP pa je tipična napoved sekundarne strukture z algoritmom 2. generacije (sekundarna strukturna stanja aminikislin: H - vijačnica α, E - list β, prazno - naključno navitje, pomišljaj pa označuje, da podatka ni). Zaradi prekratkih napovedanih listov β in omejene zanesljivosti rezultatov, se pri uporabi rezultatov TYP pojavijo naslednje težave: ne vemo ali naj 3., 4. in 5. košček (ki ni naključno navitje) v TYP združimo, ali pa program res napoveduje tri liste β (v OBS vidimo, da sta v resnici dva); v katero smer naj podaljšamo, kar želimo podaljšati; iz primerjave OBS in TYP vidimo, da so tiste vijačnice α v TYP napake (zanesljivost 60%), če imamo le TYP ne vemo ali naj jih podaljšamo ali zavržemo ipd.

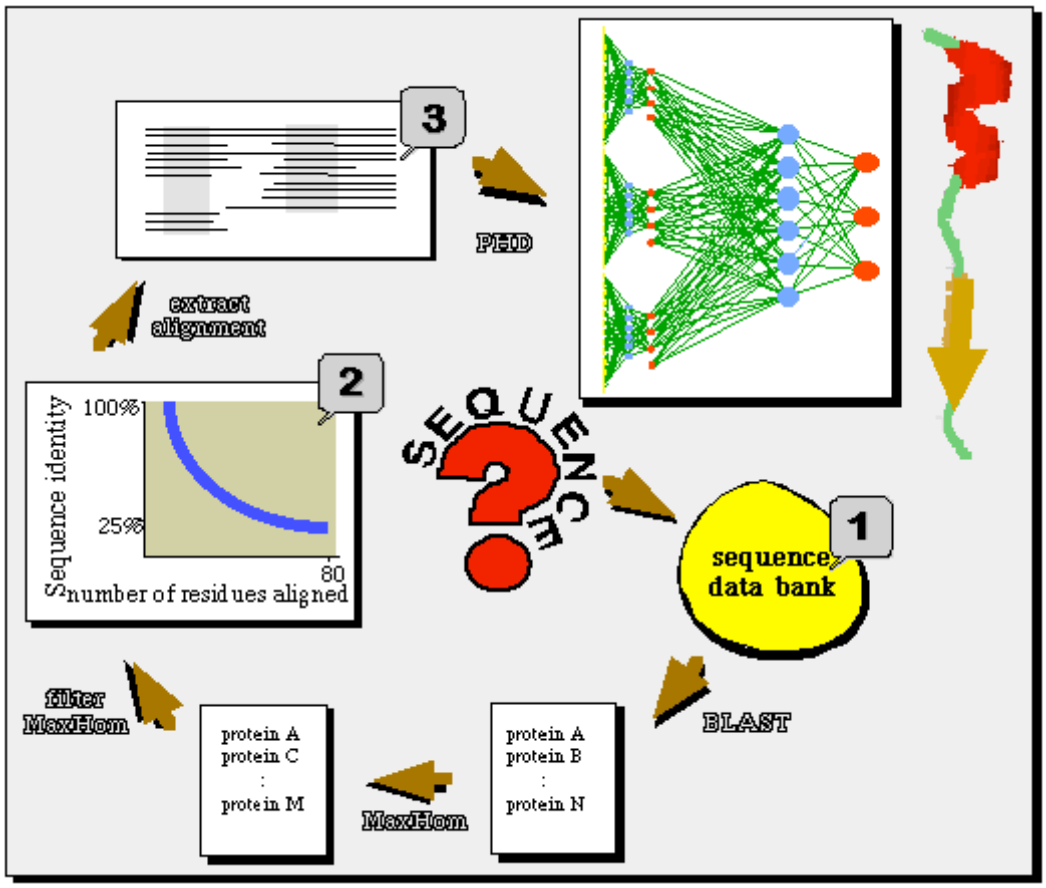
3 Algoritmi 3. generacije

Genetski material se ob prehodu iz generacije v generacijo spreminja. Prihaja do točkovnih mutacij, ko se spremeni en sam nukleotid na nekem mestu. Prihaja tudi do zdrsov - če polimeraza pri prepisovanju DNA naleti na mesto na katerem so točkovne mutacije izoblikovale ponavljajoči se vzorec, v katerem se ponavlja eden ali skupek parih nukleotidov, se z nezanemarljivo verjetnostjo lahko zgodi, da polimereza zdrsne in se ponavljajoči se vzorec podaljša ali skrajša.

Če se to dogaja v eksonih, se vse to odraža tudi na aminokislinsko zaporedje beljakovine, ki je v eksonu kodirana. Že sprememba nekaj aminokislin v beljakovini lahko beljakovinsko 3D strukturo destabilizira. Tridimenzionalna struktura beljakovin je namreč tako kompleksna, da jo lahko destabilizira ali spremeni že tudi sprememba sekundarne strukture manjšega koščka. Obstajajo pa tudi take spremembe aminokislin, ki sekundarne in 3D strukture ne spremenijo. Tridimenzionalna struktura beljakovine je tako zelo pomembna za njeno funkcijo, da vsaka pomembnejša sprememba 3D strukture spremeni tudi funkcijo beljakovine. Seveda je možno, da funkcijo izboljša, na tem tudi temelji evolucija, a to je zelo malo verjetno. Velika večina sprememb funkcijo poslabša, zato so takšni organizmi manj sposobni za preživetje in v boju za obstanek izumrejo. Če je sprememba zelo slaba, organizem seveda umre takoj. Tako torej pritisk naravne selekcije drži sekundarne in 3D strukture beljakovin, ki dobro delujejo, nespremenjene, tudi med sorodnimi ali celo manj sorodnimi vrstami. Vsi dosedaj znani pari naravnih beljakovin z znanimi 3D strukturami katerih aminokislinska zaporedja so enaka v več kot 35% aminokoslin na istoležnih mestih imajo podobne 3D strukture [4]. Še več, večina naravnih beljakovin s podobnimi 3D strukturami ima celo manj kot 15% enakih aminokislin na istoležnih mestih [4].

Pri družinah beljakovin opažamo razne substitucijske vzorce - npr. spremembe, ki ne spremenijo 3D strukture te družine beljakovin. Opaženo je bilo, da so ti vzorci so močno odvisni od 3D strukture beljakovin. Še več, substitucijski vzorci implicitno vsebujejo tudi informacijo o interakcijah med oddaljenimi aminokislinami v aminokislinskem zaporedju - npr. če sta dve aminokislini, ki sta daleč narazen v aminokislinskem zaporedju, potem, ko se beljakovina zvije, fizično dejansko ena zraven druge v prostoru, in so med njima tudi razne interakcije, potem se lahko ena od njiju zamenja le s tako aminokislino, ki po svojih fizikalno-kemijskih lastnostih prav tako ustreza oni drugi, če naj se 3D struktura ne spremeni.

Na substitucijskih vzorcih temelji *3. generacija algoritmov* za napovedovanje sekundarne strukture. Osredotočil se bom na algoritem PHD [4], ki napoveduje sekundarne strukture z 72% zanesljivostjo (slika 1). Shema



Slika 3: Shema osnovnega algoritma PHD.

osnovnega algoritma PHD je na sliki 3.

Algoritem PHD najprej aminokislinsko zaporedje beljakovine, ki ji določa sekundarno strukturo (*preučevana sekvenca*), s programom *Blast* [9] primerja z aminikislinskimi zaporedji beljakovin v eni od velikih podatkovnih baz (*bazne sekvence*) in poišče najbolj podobna. Blast poskuša najti zelo podobne kose aminokislinskih zaporedij v po dveh beljakovinah na enkrat. Ena je preučevana, ki ji iščemo podobne, druga pa je po ena iz baze (zvrstijo se vse). Poljubno lahko odloča, kje bo začel kosa, ki ju primerja, za vsako ujemanje istoležnih aminokislin je nagrajen $(+5)$, za vsako neujemanje je kaznovan (-4), proti kazni (-10) pa lahko v eno in drugo zaporedje po potrebi dela tudi luknje in s tem naredi potrebne premike za boljše ujemanje naprej po sekvencah, če je to postalo slabo, pa je tak premik smiselen. Nagrade in kazni se seštejejo, višji kot je rezultat, boljše je ujemanje. Blast poišče take

pare kosov, da je ujemanje najboljše ter jih izpiše, če so si dovolj podobni.

Blast primerja po dve aminokislinski zaporedji naenkrat, program *MaxHom* [10] pa naredi podobno prileganje večih zaporedij hkrati. To je že nasledja stopnja algoritma PHD. S programom MaxHom izvede hkratno prileganje tistih kosov baznih sekvenc, ki jih je Blast našel kot najbolj podobne preučevani, na preučevano sekvenco.

Naslednja stopnja upošteva le tiste kose, pri katerih je bilo prileganje zelo uspešno - to so homologi. Rezultat analize relacij med izbrano sekvenco in homologi ter homologi samimi predstavlja vhodne podatke za dve stopnji triplastnih (vhodna, srednja-skrita, izhodna plast) nevronskih mrež, ki predstavljajo PHD v ožjem pomenu besede. Analizira se okno 13-ih aminokislin, ki hkrati drsi po preučevani sekvenci in vzporejenih homologih.

Prva stopnja nevronskih mrež določi verjetnost za vsako od sekundarnih strukturnih stanj, da se pojavi v centralni aninokislini okna na preučevani sekvenci. Vhodni podatki za prvo stopnjo nevronskih mrež so naslednji:

1. profil aminokislinskih substitucij med preučevano sekvenco in homologi za vseh 13 aminokislinskih mest v oknu,

2. uteži izračunane na podlagi enakosti istoležnih aminokislin med homologi (evolucijsko ohranjanje aminokislin) za vsako od aminokislinskih mest v oknu,

3. v evoluciji se lahko tudi kaka aminokislina doda (zdrs + mutacija) ali odstrani (zdrs) - število vstavkov in izbrisov pri homologih glede na preučevano sekvenco za vsako mesto v oknu je tretji vhodni podatek za nevronske mreže,

4. položaj okna glede na začetek in konec preiskovane sekvence (upoštevajo se tudi luknje),

5. aminokislinska zgradba okna na preučevani sekvenci,

6. dolžina preučevane sekvence.

Če bi za sekundarno strukturno stanje posamične aminokisline vzeli kar tisto, ki ga je kot najbolj verjetno določi prva stopnja nevronskih mrež, ne bi upoštevali medsebojnih korelacij med sekundarnimi strukturnimi stanji sosednjih aminokislin. Prva stopnja nevronskih mrež je namreč učena tako, da se za preučevano sekvenco jemljejo naključno izbrani izseki raznih aminokislinskih zaporedij z znanimi sekundarnimi strukturami z dolžino enega okna (13 aminokislin), ki nimajo nobene medsebojne povezeve, v splošnem

niso niti iz iste beljakovine, kaj šele, da bi si sledili v aminokislinskem zaporedju. Zato tako ne bi bile upoštevane naravne dolžine posamičnih elementov sekundarne strukture - npr. vijačnica α mora vsebovati najmanj tri aminokisline. Torej je potrebno narediti tudi nekakšen kompromis med sosedi, da se elementi sekundarne strukture raztezajo čez več aminikislin tako kakor v naravi. To opravi druga stopnja nevronskih mrež, katere razultat je sekundarna struktura preiskovane sekvence.

Algoritem PHD ima več parov prve in druge stopnje nevronskih mrež, ki so učeni ločeno. Izvede vse, potem pa na tretji stopnji vse te rezultate izpovpreči. Tako dobimo končno napoved sekundarne strukture preučevane beljakovine.

4 Zaključek

Znane beljakovinske sekvence (aminokislinska zaporedja) so zbrane v podatkovni bazi *SWISS-PROT*. Baza je javno dostopna. Trenutno vsebuje več kot 200 000 sekvenc. Tiste med njimi, katerim so z rentgenskim sipanjem ali NMR določili tridimenzionalno strukturo, pa so zbrane v prav tako javno dostopni bazi *PDB*. Ta vsebuje le nekaj več kot 30 000 vnosov, saj je postopek eksperimentalne določitve 3D strukture zamuden. Ker se odkrivanje aminokislinskih zaporedij beljakovin nadaljuje z nezmanjšano hitrostjo, ni pričakovati, da bi se ta razkorak v doglednem času zmanjšal, zato si že dolgo prizadevajo razviti čim natančnejše algoritme za računalniško napovedovanje sekundarne, pa tudi terciarne in kvartarne strukture. Algoritmi za napovedovanje zadnjih dveh so še na začetku svojega razvoja, lepe uspehe pa so dosegli z algoritmi za napovedovanje sekundarne strukture. Na začetku, ko so določali sekundarno strukturno stanje posamične aminokisline v verigi le na podlagi statistične pogostnosti posamičnih vrst aminokislin v posamičnih elementih sekundarne strukture je bila njihova zanesljivost okrog 50%. Ko so upoštevali še vplive okoliških aminokislin na sekundarno strukturno stanje vsake aminokisline, se je zanesljivost dvignila že nad 60%. Trenutno pa so najboljši algoritmi, ki napovedo sekundarno strukturo beljakovine na podlagi podatkov o evolucijskem razvoju te beljakovine, ki jih pridobijo s primerjavo njene sekvence s sekvencami njenih homologov. Zanesljivost teh algoritmov je nad 70%.

Literatura

[1] P. Tompa, *BioEssays* **25**, 847 (2003)

[2] W.-M. Zheng, *Protein secondary structure prediction based on quintuplets*, http://xxx.lanl.gov, **arXiv:physics/0307076 v1** 16 Jul 2003

[3] X. Liou, L.-M. Zhang, W.-M. Zheng, *Prediction of protein secondary structure based on residue pairs*, http://xxx.lanl.gov, **arXiv:physics/0212065 v3** 15 Jun 2004

[4] B. Rost, *Rising accuracy of protein secondary structure prediction*, v knjigi *Protein structure determination, analysis, and modeling for drug discovery*, urednik: D. Chasman, (Dekker, New York) str. 207

[5] B. Rost, C. Sander, *3rd Generation Prediction Of Secondary Structure*, v knjigi *Predicting protein structure*, urednik: D. M. Webster, (Humana Press, 1998)

[6] J. Garnier, D. Osguthorpe in B. Robson, *J.Mol.Biol.* **120**, 97 (1978)

[7] J. F. Gibrat, J. Garnier in B. Robson, *J.Mol.Biol.* **198**, 425 (1987)

[8] J. Garnier, J. F. Gibrat in B. Robson, *Meth.Enzymol.* **266**, 540 (1996)

[9] S. F. Altschul in W. Gish, *Meth.Enzymol.* **266**, 460 (1996)

[10] C. Sander in R. Schneider, *Proteins* **9**, 56 (1991)