CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice [1]

Jakub Chromý, Zuzana Kroulíková, Róbert Eckhaus

Apríl 2015

Voľne dostupný program nazvaný *CLUSTAL W* predstavoval v roku 1994 novú multiple sequence alignment metódu umožňujúcu hľadanie spoločných vzorov sekvencií k detekcii proteínových rodín, určenie homológie, predikciu sekundárnej a terciárnej štruktúry nových sekvencií, navrhovanie primerov či štúdium evolučnej príbuznosti v spojitosti s citlivejším prístupom k dátam a vysokou efektivitou.

Multiple sequence alignment (MSA) je nepostrádateľná metóda v molekulárnej biológii. V súčasnej dobe so vzrastajúcou intenzitou sekvenovania rastú i požiadavky na zvýšenie presnosti MSA, pokiaľ možno bez negatívneho vplyvu na rýchlosť. Práve to umožnil so svojimi inováciami *CLUSTAL W*.[1]

Pri klasickom alignmente dvoch sekvencií sa používa dynamické programovanie so skórovacími tabuľkami, ktoré je však pri MSA možné využiť len pri malom množstve krátkych sekvencií – limitujúci je v tomto prípade výkon súčasných počítačov, a teda neprimerane dlhá doba na zostavenie alignmentu. [2] Preto *CLUSTAL W* (a vo všeobecnosti všetky MSA programy) pri alignmente viacerých sekvencií o väčšej dĺžke využíva heuristické algoritmy, ktoré síce vypočítajú len približne optimálny alignment, ale za prípustný čas.

Najdôležitejším predpokladom je, že homologické sekvencie sú si i evolučne blízke a teda, že MSA je možné relatívne rýchlo zostaviť sériou párových alignmentov, ktoré kopírujú vetvenie fylogenetického stromu. Pritom sa ako prvé zostavujú veľmi presné alignmenty najpríbuznejších sekvencií, v ktorých sa určia preferované pozície medzier – tie sa už pri nasledujúcich alignmentoch so vzdialenejšími sekvenciami nemenia. Práve tu môže byť potenciálny problém, keďže chyby v prvotných alignmentoch zostanú fixované a ani pridávanie ďalších informácií z neskorších porovnávaní so vzdialenejšími sekvenciami ich už nijak neovplyvní.

Táto chybovosť môže plynúť z nesprávneho zostavenia východzieho fylogenetického stromu, kedy vetvenie stromu môže byť nesprávne ovplyvnené vysokou lokálnou homológiou na úkor objektívnejšej celkovej homológie. Väčšina chýb však vzniká „jednoduchými“ misalign chybami v prvotných párových alignmentoch najpríbuznejších sekvencií. Druhým závažným problémom môže byť výber vhodných alignment parametrov.

Pri jednoduchých alignmentoch sa využívajú 2 rôzne gap penalty –

[1] *CLUSTAL W* rozširuje algoritmus progresívneho alignmentu Fenga a Doolittla [2], ktorý bol v dobe vzniku článku dominantnou metódou výpočtu MSA

[2] Optimálny MSA patrí do skupiny tzv. NP-kompletných problémov [3]. To znamená, že pravdepodobne neexistuje vôbec žiaden "rýchly" algoritmus, ktorý by ho riešil. Dnes dostupné programy pre výpočet optimálneho MSA sú zväčša založené na dynamickom programovaní a dokážu spracovať len zadania s rádovo menej ako 10 sekvenciami.

jedna pre vznik a druhá pre rozšírenie medzery. Takýto systém funguje dobre pri sekvenciách, ktoré sú od seba evolučne rovnako vzdialené. Problém nastane pri alignmente sekvencií, ktoré si sú rôzne podobné - systém, ktorý by používal jedinú skórovaciu maticu pre všetky sekvencie by zavádzal a bol by značne nepresný.

Tento problém rieši *CLUSTAL W* tým, že flexibilne mení skórovacie tabuľky a hodnoty gap penalt v závislosti na fylogenetickej vzdialenosti sekvencií. Táto inovácia výrazne zvýšila senzitivitu MSA. V ideálnom prípade získame už v tomto kroku alignment o vysokej kvalite, z ktorého sa dajú správne odvodzovať sekundárne resp. terciárne štruktúry. V zložitejších prípadoch dostaneme aspoň východzí alignment, ktorý poslúži ako dobrý základ pre ďalšie spracovanie.

Metódy programu

Beh programu prebieha v troch hlavných fázach:

1. Výpočet **matice vzdialeností** sekvencií: pre každú dvojicu sekvencií sa vypočíta ich vzdialenosť pomocou párového (*pairwise*) alignmentu. V programe *CLUSTAL W* má užívateľ na výber 2 metódy: optimálne riešenie za použitia dynamického programovania, alebo rýchly aproximačný algoritmus.

2. Na základe matice vzdialeností sa zostaví fylogenetický strom pomocou metódy zlučovania susedov. [3] Tá vracia nezakorenený strom. Koreň určíme ako miesto, v ktorom bude priemerná dĺžka vetiev oboch jeho podstromov rovnaká. Tento postup nám dáva výsledný **vodiaci strom** *(guide tree)* zodpovedajúci podobnosti jednotlivých sekvencií, kde dĺžky hrán sú priamo fylogenetické vzdialenosti a listy odpovedajú sekvenciám.

3. Sekvencie sa **progresívne zrovnávajú** v poradí určenom vypočítaným stromom. Postupuje sa od evolučne najbližších dvojíc k najvzdialenejším. Alignment sa vykonáva metódou dynamického programovania a vstupujú doňho buď listy vodiaceho stromu (jednoduché sekvencie), alebo už spracované vnútorné vrcholy (hotové paiwise alignmenty medzi sebou navzájom alebo s jednoduchými sekvenciami).

[3] Pri výpočte fylogenetických stromov sa často používa princíp maximálnej úspornosti (parsimónie). Ten preferuje stromy reprezentujúce čo najmenej evolučných zmien. Nájsť najúspornejší strom je však znovu výpočetne nezvládnuteľná úloha. *CLUSTAL W* preto opäť využíva heuristickú metódu - tzv. *Neighbor Joining* [4]. Algoritmus vychádza z počiatočnej hviezdicovej topológie stromu a postupne zhlukuje susediace taxóny tak, aby po každom kroku bol súčet dĺžok hrán stromu minimálny (pričom zlúčené taxóny sa berú už iba ako jeden list).

Váhy sekvencií

Každej sekvencii je priradená váha vyjadrujúca jej relatívnu dôležitosť. Tento krok reaguje na situáciu, v ktorej by mohlo viacero príbuzných sekvencií "zatieniť"jediný exemplár evolučne vzdialenejšej sekvencie, ktorá poskytuje systému novú informáciu. Preto je príbuzným sekvenciám priradená nižšia váha (môžeme o nich principálne uvažovať ako

o jednej skupine, ktorej dôležitosť sa rozdelí medzi všetkých členov).

Samotný alignment

Pri samotnom progresívnom alignmente sa už vždy využíva dynamické programovanie. Postup pri skórovaní dvojíc residuí však musel byť pozmenený vzhľadom k tomu, že nepracujeme so sekvenciami, ale s celými alignmentami sekvencií (jednoduchú sekvenciu môžeme považovať za špeciálny prípad alignmentu).

Vezmime si teda 2 alignmenty a a b pozostávajúce z n resp. m sekvencií $(a_1 \dots a_n)$, $(b_1 \dots b_m)$. Prvok na pozícii n sekvencie s označme $s(n)$. Potom nevážené skóre $S'_{a,b}(x,y)$ alignmentov a, b na pozíciách x, y so skórovacou maticou M vypočítame ako priemer skóre všetkých dvojíc residuí na daných pozíciách.

Algoritmus však uvažuje aj vyššie popísané relatívne dôležitosti sekvencií. Príspevky jednotlivých dvojíc residuí v $S'_{a,b}(x,y)$ preto ešte prenásobíme váhami w sekvencií, z ktorých pochádzajú. Výsledné skóre bude teda férovo zohľadňovať aj znaky z odľahlejších a "podreprezentovaných"sekvencií fylogenetického stromu. Dostávame výsledné skóre:

$$S_{a,b}(x,y) = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} (w_i \cdot w_j \cdot M(a_i(x), b_j(y)))$$

Gap penalties

Ďalším prínosom *CLUSTAL*-u je dynamický charakter tzv. *Gap penalties*, teda postihov udeľovaných pri vytvorení alebo rozšírení medzier. Rozlišujeme *Gap Opening Penalty (GEP)* - postih za vytvorenie novej medzery a *Gap Extension Penalty (GEP)* - postih za rozšírenie medzery o jedno residuum. Počiatočné hodnoty sú nastavené užívateľom a algoritmus sa ich v priebehu výpočtu snaží upravovať tak, aby zodpovedali charakteru dát, s ktorými pracuje.

Gap Opening Penalty závisí od

1. skórovacej tabuľky: prepočítava sa podľa priemerného skóre misaligned dvojice.
2. podobnosti sekvencií: zvyšuje sa lineárne so zvyšujúcou sa podobnosťou porovnávaných sekvencií. Takto dostaneme konzervatívnejšie alignmenty u podobných sekvencií (väčšiu preferenciu identít).
3. dĺžky sekvencií: zvyšuje sa so zvyšujúcou sa dĺžkou sekvencií

Gap Extension Penalty je závislá na rozdiele dĺžok sekvencií a bude nastavná na vyššiu hodnotu, ak je jedna zo sekvencií oveľa kratšia ako tá druhá.

Gap Penalties závislé od pozície v sekvencii

CLUSTAL W zavádza tiež modifikácie *gap penalties* na základe lokálnych vlastností sekvencie. Pred každým alignmentom sa vypočíta pre každú pozíciu v oboch sekvenciách pravdepodobnosť, že sa tam nachádza začiatok medzery. Lokálne hodnoty *GAP* a *GOP* ovplyvňujú:

1. Ak je na danej pozícii medzera, tak tu znížime *GOP*. Znížená hodnota *GOP* sa pohybuje od 30% pôvodnej hodnoty (ak sa jednalo o jedinú medzeru v alignmente mnohých sekvencií) po 0% (ak je medzera vo všetkých sekvenciách alignmentu). *GEP* sa vždy zníži na polovicu.
2. Medzery sa zväčša nevyskytujú príliš blízko seba. Ak na danej pozícii nie je žiadna medzera, ale v okruhu 8 residuí nejaká je, zvýšime *GOP* zhruba dvoj- až štvornásobne (podľa vzdialenosti od najbližšej medzery).
3. Postupnosť 5 a viac hydrofilných residuí nazývame hydrofilný úsek (*hydrophilic stretch*). V prípade, že sa na pozícii nenachádza medzera a je sučasťou hydrofilného úseku, znížime tu GOP o tretinu.
4. Experimentálne boli zistené frekvencie výskytu medzier po jednostlivých residuách. Po niektorých residuách je otvorenie medzery pravdepodobnejšie. V prípade, že skúmaná pozícia nebsahuje v žiadnej sekvencii medzeru ani neleží na hydrofilnom úseku, prenásobime GOP váhou [4] podľa tabuľky frekvencií výskytu medzier.

[4] Ak sa jedná o alignment, vypočítame priemer cez váhy residuí jednotlivých sekvencií

Skórovacie tabuľky

V *CLUSTAL W* sú k dispozícii 2 triedy tabuliek - BLOSUM (predvolená) a PAM. Každá trieda obsahuje tabuľky špecializované na rôzne príbuzné sekvencie. Príbuznosť sa zistí priamo z pomocného stromu a na základe nej sa vyberie najvhodnejšia tabuľka.

Odloženie výpočtu vzdialených sekvencií

Niektoré sekvencie môžu byť silne divergujúce od zvyšku dát a preto sa s nimi počíta len veľmi obtiažne. Informácie nazbierané počas behu MSA nám však vedia napovedať, kde sú pravdepodobnejšie miesta začiatku medzier a ako zarovnať slabo zakonzervované časti sekvencie.

Preto *CLUSTAL W* poskytuje možnosť odložiť výpočet alignmentov divergentných sekvencií až na koniec behu programu, keď už budú tieto pomocné informácie dostupné. Za divergujúcu pritom považujeme sekvenciu, ktorá ma sekvenčnú identitu nižšiu ako ísta medzná hodnota (napr. 40%) so všetkými ostatnými sekvenciami v alignmente.

Porovnanie s ďalšími metódami

V dobe vzniku *CLUSTAL W* bolo hlavným problémom nastavenie pozične-špecifických parametrov pre multiple alignment. Ostatné programy sa tento problém snažili riešiť buď prostredníctvom navýšenia lokálných gap penalt v pravidelných sekundárnych štruktúrach, pričom vychádzali z už známych 3D štruktúr; alebo v druhom prípade prostredníctvom tzv. skrytého Markovho modelu, ktorým sa určili pozične-špecifické gap penalty a váhy jednotlivých residuálnych substitúcií, čo ale vyžadovalo veľký počet už známych proteínových domén. *CLUSTAL W* však tieto informácie získaká už priamo zo súboru porovnávaných sekvencií, preto mohol byť aplikovaný na ľubovoľný set sekvencií a to i bez známej vyššej štruktúry. Úspešnosť tejto metódy závisí na množstve sekvencií, ich evolučnej príbuznosti, na presnosti maticových parametrov a na schopnosti vhodne prispôsobovať tieto parametry v priebehu alignmentu.

Záver a diskusia

Program *CLUSTAL W* výrazne zlepšil senzitivitu bežne používaných multiple sequence alignment metód pre alignment rôznych proteínových sekvencií.

Inovácie tento pokrok umožňujúce boli: za prvé, ku každej sekvecii je priradená konkrétna váha v čiastočnom alignmente za účelom zníženia dôležitosti blízko príbuzných sekvencií a vyzdvyhnutia tých naviac vzdialených. Za druhé, substitučné matice pre aminokyseliny sú striedané pri rôznych alignmentoch v závislosti na príbuznosti porovnávaných sekvencií. Za tretie, gap penalties závislé od výskytu konkrétnej aminokyseliny na danej pozícii a miestne redukované gap penalties v hydrofilných oblastiach vedú ku vzniku nových medzier skôr v potenciálnych loop oblastiach ako v pravidelnej sekundárnej štruktúre. Za štvrté, polohy, kde boli v skorších alignmenoch vytvorené medzery, dostávajú redukovanú lokálnu GOP, čo podporuje tvorbu nových medzier na týchto pozíciách.

Výhodou programu *CLUSTAL W* bola jeho schopnosť pracovať s väčším množstvom dát, než bolo do tej doby možné, naviac s väčšou presnosťou, rýchlosťou a celkovou efektivitou. Prielomovým bol fakt, že bola umožnená práca s akýmikoľvek skupinami sekvencií bez

ohľadu na to, či je z danej množiny sekvencií nejaká štruktúra už známa.

Rozvoj v získavaní dát o proteínových doménach či proteínových rodinách neustále rastie, čo mimo prísun nových potenciálne využiteľných dát vedie i k väčšej presnosti programov a porovnávania štruktúr. V súčasnosti je už *CLUSTAL W* nahradený programom *CLUSTAL Omega* [5], ktorý je ešte výkonnejší, teda o mnoho rýchlejší a schopný porovnávať aj státisíce sekvencií naraz.

Literatúra

[1] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, Nov 1994.

[2] D. F. Feng and R. F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, 25(4):351–360, 1987.

[3] W. Just. Computational complexity of multiple sequence alignment with SP-score. *J. Comput. Biol.*, 8(6):615–623, 2001.

[4] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, Jul 1987.

[5] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7(1), 2011.