Testo originale in russo
http://zbio.net/bio/001/003.html

Traduzione a cura di M.T. Carreras
http://geneproject.altervista.org/

Analisi filogenetica delle famiglie di proteine omologhe

Naumoff D.G.

Laboratory of Bioinformatics,
State Institute for Genetics and
Selection of Industrial Microorganisms,
Moscow 117545, Russia

http://bioinform.genetika.ru/members/Naumoff/index.htm

ricevuto - 27 gennaio 2006;
accettato - 22 febbraio 2006;
pubblicato - 4 aprile 2006

Nell'articolo viene descritta la metodica di esecuzione di un'analisi filogenetica di famiglie di proteine. Si assume che "in ingresso" si abbia la sequenza aminoacidica della proteina, mentre "in uscita" si richiede:

i. ottenere un albero filogenetico della corrispondente famiglia di proteine;
ii. determinare all'interno di questa famiglia le sottofamiglie;
iii. riesaminare le relazioni evolutive della famiglia studiata con altre famiglie.

I metodi operativi proposti sono applicabili a qualunque proteina, ma operano con effetto ottimale su proteine globulari solubili. È possibile analizzare un fermento, avente una nota sequenza aminoacidica, che è stato da voi personalmente caratterizzato biochimicamente. In pratica per tale analisi passerà anche l'ipotetica proteina ottenuta dai database, codificata come Open Reading Frame in un certo genoma sequenziato recentemente. La relazione è illustrata con esempi ottenuti dall'analisi di diverse famiglie di glicoside idrolase. I programmi per computer utilizzati nell'articolo sono gratuiti e ottenibili via internet.

1. Oggetto di analisi

Proteine-omologhe – gruppo di proteine di uno e/o diversi organismi, i cui geni nella maggior parte dei casi hanno una comune origine evolutiva. Le cause della comparsa delle proteine-omologhe possono essere diverse: divergenza degli organismi (trasmissione verticale), duplicazione dei geni e dei genomi, trasmissione orizzontale.

Con il termine *famiglie di proteine-omologhe* 15-20 anni fa poteva essere

intesa tutta l'aggregazione delle proteine tra di loro omologhe. Tuttavia, l'accresciuta sensibilità nei metodi di confronto tra le sequenze aminoacidiche e il rapido accumulo dei dati delle strutture terziarie di proteine più conservative, ha rivelato la parentela evolutiva tra molte delle famiglie precedentemente conosciute. Il termine "famiglia" è ora consunto e molti autori lo possono trattare in modo differente. Spesso l'appartenenza di una proteina ad una particolare famiglia implica la presenza di una specifica, nota o presunta, attività enzimatica (oppure di un'altra funzione biologica), proprio in base alla quale si dà il nome alla famiglia. Tuttavia, tale ambiguità spesso porta al caso di due distinte famiglie descritte dal medesimo autore che possono essere viste da altri ricercatori come una sola famiglia. Dal punto di vista dell'analisi filogenetica ciò non ha alcun significato intrinseco. Quello che ha più importanza è che le proteine della stessa famiglia abbiano formato un gruppo monofilogenetico, ed il grado di somiglianza delle loro sequenze aminoacidiche sia abbastanza elevato da permettere la costruzione di un allineamento multiplo globale. Attualmente il numero di famiglie di proteine note è di circa 10 mila, per esempio nel database Pfam è presente una lista di 8183 famiglie.

Ancora un altro problema relativo alla distinzione tra le famiglie risulta essere la complessa struttura di dominio di molte proteine. Le *strutture di dominio* delle proteine vengono meglio distinte attraverso l'analisi della loro struttura spaziale. L'esistenza di dati sperimentali di strutture tridimensionali permette di individuare il numero dei domini e le zone di confine fra di loro nella struttura primaria della proteina. Come è solito, domini di struttura diversi eseguono funzioni biologiche diverse, rivelandosi essere proprio i *domini funzionali*. L'assenza di informazione sulla struttura spaziale della proteina complica considerevolmente l'identificazione della sua struttura di dominio. Spesso domini diversi della stessa proteina hanno una storia evolutiva indipendente. In questo caso, essi risultano essere anche *domini evolutivi*. Tuttavia, in molti casi, due domini di struttura sono presenti quasi sempre contemporaneamente nella proteina dando origine ad un unico dominio evolutivo. Per esempio, tali strutture di domini appaiati le posseggono le famiglie glicoside-idrolase GH27 (Fig. 1) e GH32.

È comunemente accettato suddividere le grandi famiglie di proteine omologhe in *sottofamiglie*, sulla base del confronto dei livelli di similarità delle loro sequenze aminoacidiche; tuttavia, regole comuni speciali non esistono. Le famiglie evolutivamente imparentate vengono spesso riunificate in *super famiglie* (oppure *clan*). Per esempio, nel database Pfam, in 206 clan sono raggruppate 1396 famiglie.

2. Spiegazione della struttura di dominio della proteina e scelta del dominio di analisi

Per questo lavoro è necessario scegliere una proteina con una sequenza aminoacidica nota. Di regola, la molecola proteica è costituita da alcune centinaia di residui aminoacidici. È possibile che la proteina in esame (o taluni dei suoi omologhi) sia costituita da diversi domini strutturali. Tale possibilità cresce rapidamente con l'aumentare della lunghezza della sequenza aminoacidica (solitamente, un dominio non è mai più di 300 aminoacidi).

A seconda del problema da risolvere può essere necessario:

i. esaminare la filogenia di uno dei domini della proteina in esame, per esempio quello catalitico;
ii. esaminare tutti i suoi domini;
iii. esaminare tutta la varietà dei domini che si incontrano nei rappresentanti della corrispondente famiglia di proteine omologhe.

L'analisi filogenetica di ciascun dominio dev'essere condotta in modo indipendente utilizzando gli allineamenti multipli corrispondenti. Se due qualsiasi domini strutturali formano un dominio evolutivo comune, allora i loro alberi filogenetici devono avere una topologia similare. In questa situazione ha senso costruire un albero comune dell'intero dominio evolutivo. In altri casi, più

probabilmente, la storia evolutiva di svariati domini si diversificherà significativamente e il confronto dei loro alberi filogenetici aiuterà ad illustrare questo. È necessario notare che, la costruzione di alberi filogenetici è possibile solo in presenza di non meno di quattro rappresentanti proteici (domini) della famiglia analizzata.

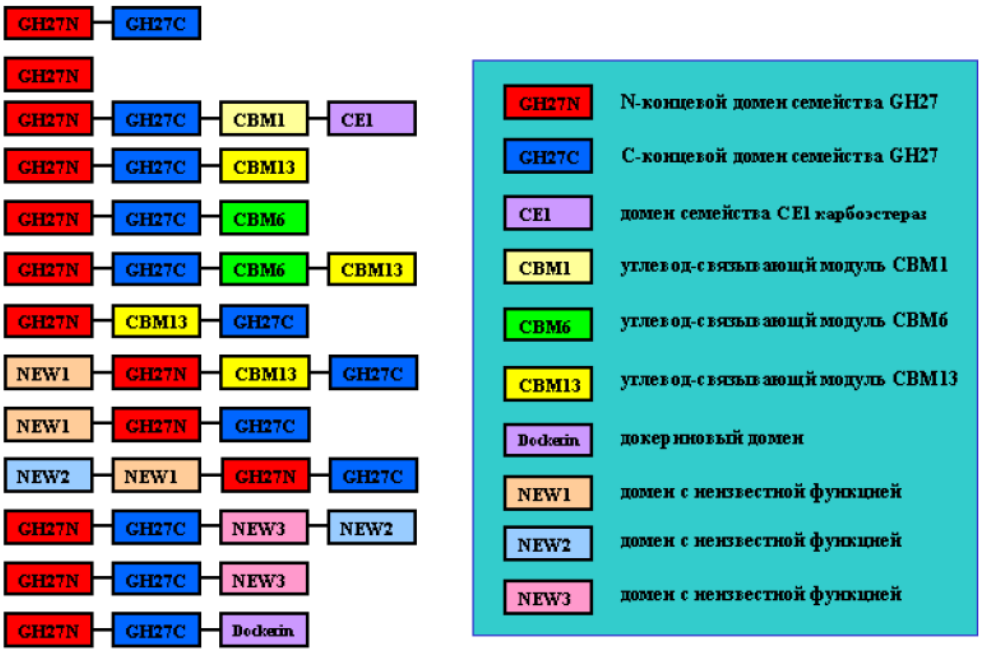

Fig. 1. Struttura di dominio delle proteine della famiglia glicoside-idrolase GH27 (1). La maggior parte delle proteine di questa famiglia sono costituite dai due domini: GH27N e GH27C. Solo qualche proteina contiene esclusivamente il dominio catalitico GH27N. Alcune proteine contengono anche dei domini aggiuntivi di tipi diversi.

Nel caso in cui la struttura di dominio della proteina analizzata non sia nota in partenza può essere determinata approssimativamente con l'aiuto di un semplice screening del database di sequenze aminoacidiche del programma blastp. Se, da un allineamento a due tra la sequenza aminoacidica della proteina ed i suoi omologhi, emerge che alcuni frammenti della proteina in esame mostrano una similarità con svariate proteine, ed i confini tra questi frammenti possono essere determinati in modo abbastanza chiaro, allora ciascuno di questi frammenti può essere visto come un dominio evolutivo a sé stante.

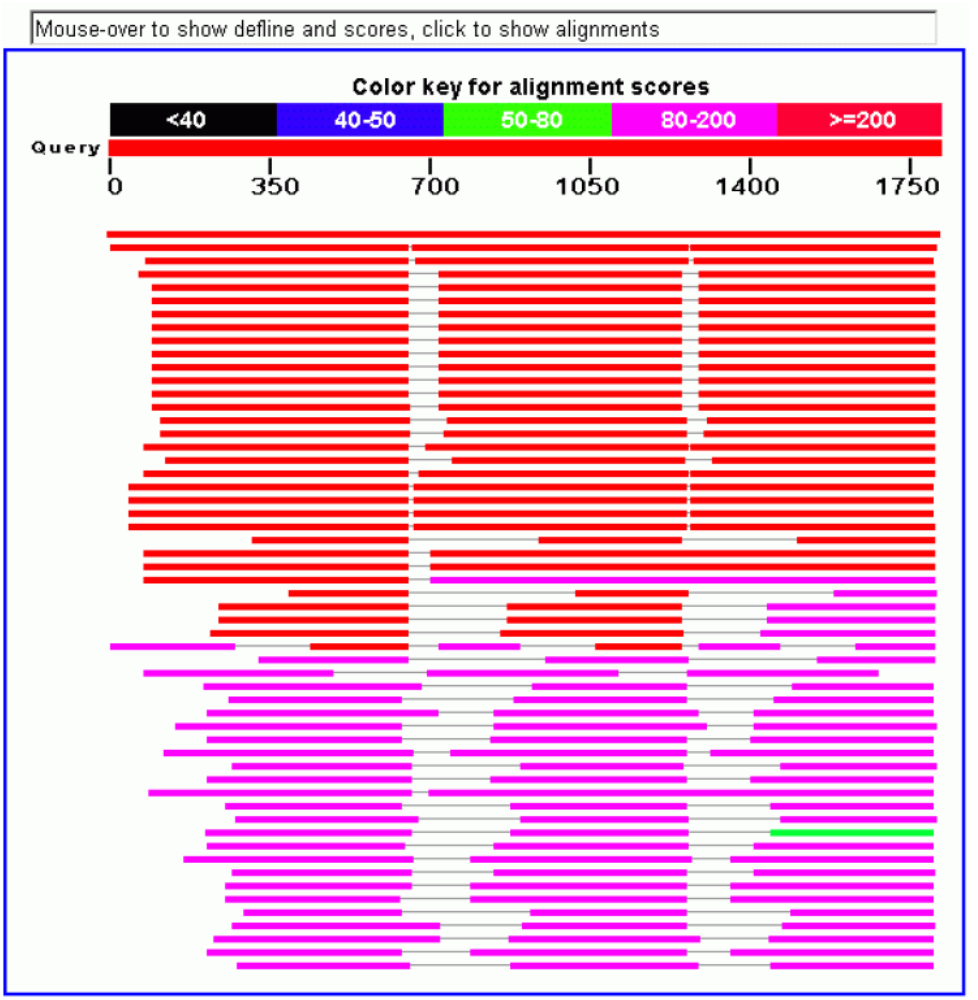

Fig. 2. Schema dimostrativo dei risultati di una ricerca di omologhi mediante il programma PSI-BLAST. Per la ricerca viene utilizzata una proteina formata da tre domini omologhi fra di loro.

3. Composizione dell'elenco di proteine della famiglia in esame

E' necessario a questo punto trovare la massima quantità possibile di proteine contenenti domini omologhi al dominio analizzato. Cioè comporre un elenco completo di rappresentanti della famiglia in esame. È possibile che la proteine in esame appartenga ad una ben nota famiglia. In questo caso ci sono buone probabilità di trovare su internet un elenco ben annotato dei suoi rappresentanti. Per esempio, nel caso di glicoside-idrolase, sul sito CAZy esiste una dettagliata classificazione di questi fermenti costruita sulla base dell'omologia ed un elenco regolarmente aggiornato dei rappresentanti di ciascuna delle centinaia famiglie glicoside-idrolase. In questo caso, è sufficiente rinnovare l'elenco di proteine della corrispondente famiglia ed aggiungervi i rappresentanti mancanti. Prevalentemente si tratterà di proteine recentemente apparse nei database che non sono ancora state associate a famiglie concrete. Se la proteina in esame appartiene ad una famiglia poco conosciuta allora l'elenco dei suoi omologhi può essere trovato per tentativi in una delle tante classificazioni globali delle proteine, tra le quali possono servire come esempio i database:

- Pfam: http://pfam.wustl.edu/
- COG/KOG: http://www.ncbi.nlm.nih.gov/COG/
- InterPro: http://www.ebi.ac.uk/interpro/
- PUMA2: http://compbio.mcs.anl.gov/puma2/cgi-bin/index.cgi
- HOMSTRAD: http://www-cryst.bioc.cam.ac.uk/homstrad/
- SCOP: http://scop.mrc-lmb.cam.ac.uk/scop/
- CATH: http://www.biochem.ucl.ac.uk/bsm/cath/cath.html

- PIR: http://pir.georgetown.edu/

E' possibile che una proteina in esame appartenga ad una famiglia non ancora caratterizzata; in questo caso l'elenco dei suoi omologhi dovrà essere composto autonomamente partendo "da zero".

In ogni caso, che si riesca a trovare un elenco di proteine abbastanza completo della famiglia in esame oppure no e si debba cominciare con l'unico rappresentante, è necessario eseguire lo screening di uno o più database con l'ausilio dei programmi della famiglia blast (Basic Local Alignment Search Tool). In modo più semplice è possibile cavarsela con il programma PSI-BLAST (Position-Specific Iterated BLAST), il quale permette, con il risultato della sua prima iterazione, di trovare nel database di sequenze aminoacidiche GenPept un elenco proteico abbastanza completo della data famiglia. In pratica un elenco esauriente di famiglie può essere ottenuto dopo la seconda iterazione (in aggiunta emergeranno i rappresentanti più divergenti della famiglia). In questo caso, un problema significativo può presentarsi nella distinzione dei rappresentanti della famiglia in esame dalle proteine-omologhe di altre famiglie. La pratica evidenzia che i rappresentanti più prossimi ad una data famiglia possono essere considerate quelle proteine ottenute dal risultato della prima iterazione. Per la composizione di un elenco più completo, ha senso, in termini di domanda (query) nello screening dei database, utilizzare diverse proteine della famiglia in esame (preferibilmente le più divergenti). Proprio per questo è molto utile avere inizialmente almeno un elenco incompleto di rappresentanti della data famiglia. Può essere utilizzato l'elenco di proteine ricavate dalla prima iterazione. E' importante non includere erroneamente nella famiglia in esame proteine con un livello di similarità troppo basso con la proteina iniziale, utilizzata come "query". Di conseguenza, in questo caso, occorre utilizzare un livello statistico di soglia considerevolmente rigido per la distinzione tra "propri-estranei" (per esempio, E-value < 10^{-5}).

Se è necessario trovare un numero aggiuntivo di rappresentanti della famiglia analizzata ha senso condurre uno screening dei database di sequenze nucleotidiche GenBank, per mezzo del programma tblastn, e Genomic BLAST. È probabile che, per il momento, alcuni dei geni codificanti le proteine della famiglia in esame non siano stati individuati e annotati tra le già note sequenze nucleotidiche (questo riguarda innanzitutto progetti genomici non completati). Tuttavia, bisogna considerare con cautela le sequenze ottenute in tal modo, in quanto possono contenere una percentuale significativamente alta di errori, e alcune di loro possono appartenere a pseudogeni.

4. Precisazione dell'elenco di proteine della famiglia in esame

Fra le proteine trovate nella famiglia in esame possono esservi anche proteine estranee. Una delle cause è la sovra-valutazione statistica della similarità con le sequenze aventi la sostanza aminoacidica degenerata – questo problema può essere parzialmente risolto mediante l'utilizzo del filtro speciale "low complexity" durante lo screening, per mezzo dei programmi della famiglia blast. Un'attenzione particolare meritano le proteine che hanno il più basso livello di similarità con gli altri membri della famiglia, ed anche quelle che riescono ad allineare il settore corrispondente ad almeno una parte del dominio in esame. Un buon test di appartenenza alla data famiglia per una qualsiasi concreta proteina (dominio) è quello che la vede impiegata come domanda durante lo screening dei database di sequenze aminoacidiche per mezzo del programma blastp. Tutti i migliori risultati statistici dello screening devono riferirsi alle proteine di questa famiglia. Diversamente, ha senso escludere dalle successive osservazioni la proteina in esame.

Tra le proteine individuate nella famiglia in esame quasi sicuramente vi sono proteine molto simili. Per esempio copie della stessa proteina in diverse colture dello stesso tipo di batterio oppure di varianti alleliche. Per l'analisi filogenetica della famiglia questi doppioni non presentano valore ed è preferibile in questa fase rimuoverli. Tuttavia, occorre ricordare che in un genoma possono essere codificati alcuni paraloghi, sequenza aminoacidiche che si differenziano

significativamente. Per questo è proibita la rimozione formale delle proteine che fanno riferimento ad un organismo già presentato nell'elenco della famiglia. Come criterio per la rimozione di proteine molto vicine in termini di sequenze aminoacidiche può servire la soglia di identità dal 95% in su.

5. Allineamento multiplo delle proteine

L'allineamento multiplo delle proteine (domini) di una famiglia può essere eseguito in modo automatico, per esempio mediante il programma ClustalW. Tuttavia questo allineamento sarà più vicino all'optimum solo in presenza di un alto livello di identità di tutte le sequenze analizzate (più del 50%) e dall'assenza in esse di una significativa quantità di inserzioni/cancellazioni. Nei casi seguenti:

- con un grado di allineamento inferiore al 30%,
- in presenza di inserzioni estese,
- in presenza dei settori facoltativi con N-terminale

gli allineamenti automatici ottenuti non sono adatti ad una analisi filogenetica corretta e l'allineamento deve essere fatto (o redatto) manualmente. In qualità di programma-redattore adatto a tale scopo può essere raccomandato BioEdit. In questo caso, ha senso utilizzare come base gli allineamenti a due o quelli multipli ottenuti automaticamente.

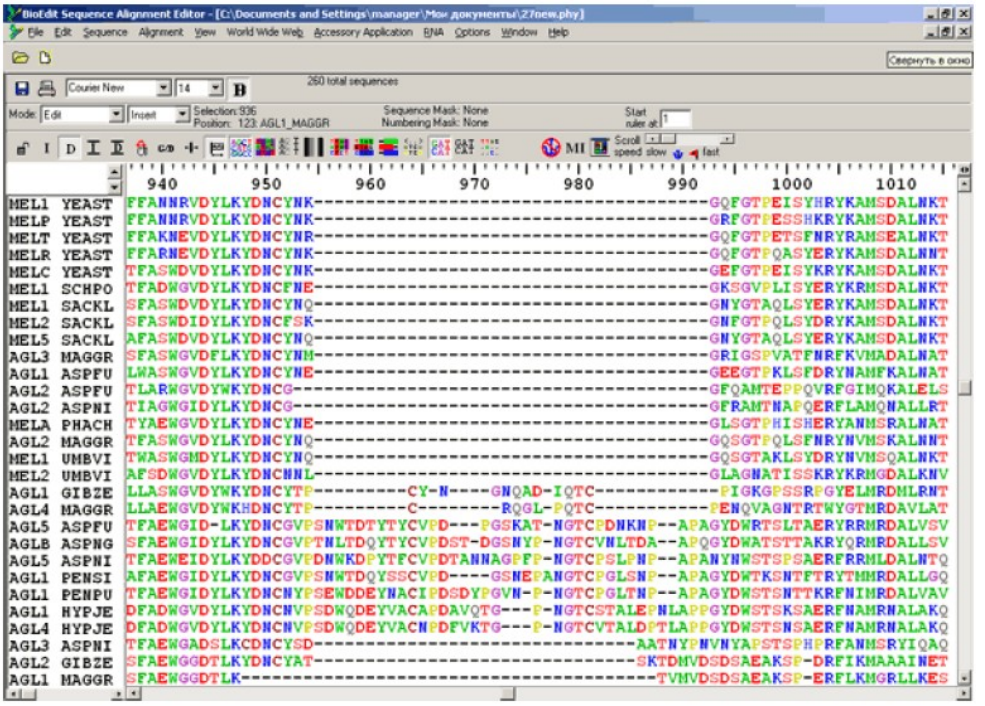

Fig. 3. Frammento dell'allineamento multiplo di sequenze aminoacidiche nel programma BioEdit.

Dopo l'ottenimento dell'allineamento multiplo ha senso osservarlo attentamente. Un'attenzione particolare deve essere rivolta a quelle proteine aventi regioni anomale nella propria sequenza:

- cancellazioni esclusive solo alla data proteina,
- basso livello di identità locale con le altre proteine in una regione per loro altamente conservativa,
- significative differenze locali tra due proteine molto simili.

E' necessaria una spiegazione delle cause che formano tali regioni anomale. Possono esservi errori di sequenziamento (per esempio, locali slittamenti del quadro di lettura), errate previsioni della struttura esone-introne, ecc. Gli errori individuati devono essere rimossi oppure le sequenze corrispondenti devono essere del tutto escluse dall'analisi successiva. Ciò riguarda anche quelle proteine che non hanno un dominio completamente analizzato (frammento della sequenza proteica).

Come ultima fase, dall'allineamento multiplo ottenuto devono essere rimosse quelle posizioni (colonne di aminoacidi) che nella maggioranza delle proteine corrispondono a cancellazione, ma anche le posizioni più incostanti la cui giustezza (univocità) nell'allineamento desta perplessità.

6. Costruzione degli alberi filogenetici della famiglia proteica in esame

L'allineamento multiplo ottenuto può essere utilizzato per la costruzione di alberi filogenetici. A questo scopo noi raccomandiamo impiegare, per esempio, i programmi PROTPARS (Protein Sequence Parsimony method) e NEIGHBOR (Neighbor-Joining method) del pacchetto PHYLIP, che permettono di eseguire un'analisi con bootstrap. È conveniente utilizzare come minimo due diversi algoritmi per la costruzione degli alberi della stessa famiglia di proteine. In questo modo le caratteristiche topologiche comuni ad entrambi gli alberi risulteranno essere affidabili criteri per la deduzione delle interrelazioni filogenetiche tra le corrispondenti proteine. L'analisi con bootstrap permetterà di valutare l'affidabilità statistica di ciascuno dei nodi dell'albero costruito. Per delle deduzioni preliminari è più che sufficiente ottenere 100 pseudorepliche di ciascun albero, sebbene le pubblicazioni scientifiche tendono ad illustrare alberi aventi 1000 pseudorepliche. Piccoli cambiamenti sul set delle sequenze e/o sul numero delle posizioni dell'allineamento multiplo permetteranno ulteriormente di verificare la stabilità degli alberi. Il programma TreeView permette di ottenere rappresentazioni grafiche degli alberi costruiti.

7. Determinazione delle sottofamiglie

Molte famiglie proteiche si rivelano essere abbastanza numerose e fra loro emergono spesso proteine che compiono differenti funzioni biologiche, per esempio fermenti con differenti attività biochimiche. Questo non permette tuttavia di predire sperimentalmente il ruolo delle proteine non ancora analizzate basandosi sui dati noti ottenuti da altri membri della data famiglia. Questo problema può essere in parte risolto tramite la suddivisione delle famiglie proteine-omologhe in sottofamiglie che uniscono evolutivamente le proteine più simili.

Una suddivisione preliminare delle famiglie proteiche in sottofamiglie può essere eseguita sulla base del confronto a due delle sequenze. Con questo deve essere scelto il più basso livello di identità delle sequenze aminoacidiche (in percentuale) che converrà alle proteine di una sottofamiglia entro i confini della famiglia analizzata. L'adeguatezza del livello scelto si avvalora tramite lo screening dei database di sequenze aminoacidiche, utilizzando diversi rappresentanti di una sottofamiglia in qualità di domanda (query). In ogni caso, solo le proteine della sottofamiglia stabilita devono avere il miglior valore della convalida statistica di similarità (E-value). Cioè, tutte loro devono rientrare nell'elenco dei risultati del programma blastp prima delle proteine delle altre sottofamiglie della data famiglia. Se non si rileva un tale scenario, è opportuno rivedere il livello di soglia di relazione tra le proteine della data famiglia verso un'altra sottofamiglia. I risultati ottenuti con l'analisi di un intero gruppo della famiglia glicosidica, hanno evidenziato che per loro un adeguato valore risulta essere dell'ordine del 30% di identità delle sequenze aminoacidiche. Tuttavia, per altre famiglie proteiche questo valore può essere del tutto differente.

Considerazioni conclusive sull'esattezza della proposta suddivisione in sottofamiglie della data famiglia proteica devono essere effettuate sulla base dei dati ottenuti dall'analisi filogenetica. In un caso ideale, tutte le sottofamiglie devono formare distinti cluster di rami sull'albero, cioè distinguersi in gruppi monofilogenetici (in qualità di gruppo esterno, durante la scelta di una qualunque sottofamiglia). I risultati di un'analisi filogenetica eseguita sulla base del confronto a due delle sequenze aminoacidiche, possono precisare la preliminare suddivisione delle famiglie in sottofamiglie, con un accurato esame sulla posizione dei singoli

rappresentanti “atipici”. Ha senso estrarre le singole sottofamiglie quando si è in presenza di almeno due rappresentanti noti. Le proteine isolate, aventi un livello di identità inferiore alla soglia rispetto a tutti gli altri rappresentati della data famiglia, devono essere osservate “per il momento” come non appartenenti a nessuna delle sottofamiglie note, in quanto le loro sequenze possono contenere errori (per esempio, locali slittamenti del quadro di lettura), che porta ad un ribassato livello di identità con le altre sequenze.

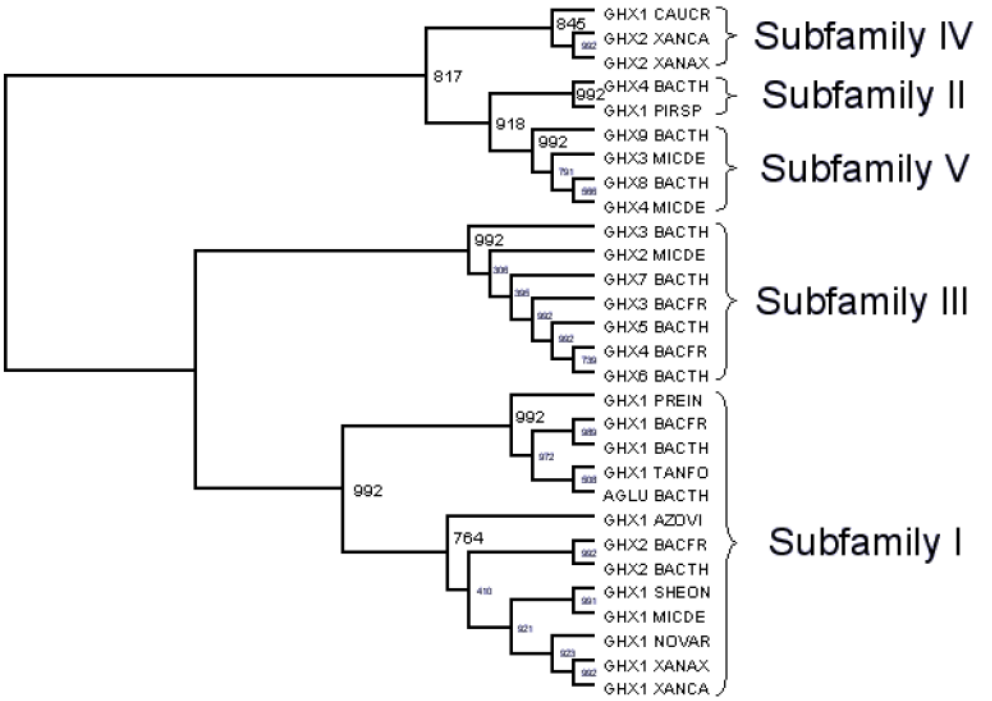

Fig. 4. L'albero filogenetico della famiglia glicosidica GH97 conferma la validità della proposta suddivisione in cinque sottofamiglie. La figura illustra i risultati pubblicati nel lavoro [2].

8. Ricerca di famiglie imparentate

Spesso risulta che nella costituzione di una qualche famiglia non c'è neanche una proteina dettagliatamente esaminata. In questa situazione, deduzioni sicure sulla struttura e sulle funzioni delle proteine di questa famiglia possono essere ottenute basandosi sulle informazioni delle proteine appartenenti a famiglie evolutivamente imparentate. Per esempio, la presenza di dati sperimentali sulla struttura terziaria di una qualche proteina permette di predire la costituzione spaziale non solo di altre proteine della stessa famiglia ma anche dei rappresentanti delle famiglie imparentate.

Per ricercare famiglie proteiche evolutivamente imparentate è opportuno utilizzare il programma PSI-BLAST. Come risultato della sua prima iterazione, solitamente, si trovano quasi esclusivamente proteine della data famiglia, mentre le successive iterazioni rivelano i rappresentanti delle famiglie imparentate. L'E-value, inteso nel senso di soglia per l'inclusione delle sequenze nella successiva iterazione, ha senso utilizzarlo a 0.01 oppure 0.001. E' necessario condurre le iterazioni fino a che non compaiono più nuove proteine con un livello di similarità assegnato. Le proteine trovate in ciascuna iterazione, devono essere esaminate per stabilire la loro appartenenza a famiglie già note oppure a delle nuove. Per questo occorre prendere in considerazione il fatto che le proteine possono contenere più di un dominio, come anche la possibilità della loro comparsa in mezzo ai risultati dello screening dei database di sequenze aminoacidiche e alle proteine non omologhe. Ne consegue che, la parentela di due famiglie di proteine è reciproca, cioè, se l'impiego delle sequenze proteiche di una famiglia permette di trovare come omologhi membri della seconda famiglia, allora l'impiego dei rappresentanti della seconda famiglia deve individuare le proteine della prima.

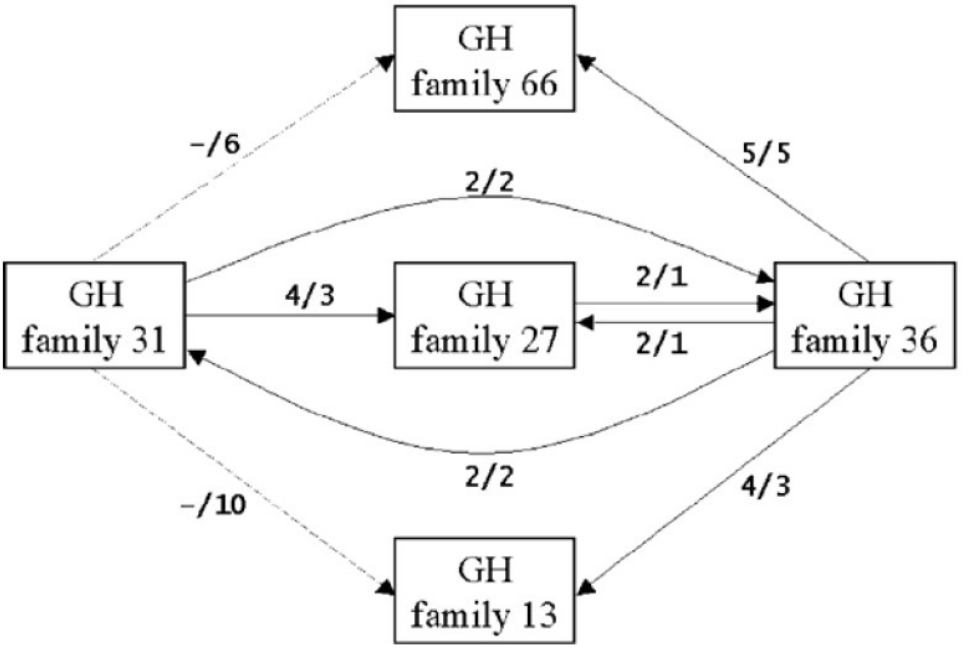

Fig. 5. Rappresentazione schematica delle relazioni evolutive determinate da PSI-BLAST. Una freccia dalla famiglia A verso la famiglia B indica che l'analisi PSI-BLAST della famiglia A ha prodotto membri della famiglia B con dei risultati significativi. I numeri x/y associati con ciascuna freccia sono i numeri di iterazione richiesti per dimostrare ciascuna relazione usando soglie E-value pari 0.001 (x) o 0.01 (y). Un trattino al posto della x indica che la relazione non si manifesta ad una tale ristretta soglia di E-value e queste deboli relazioni sono mostrate con delle linee punteggiate. Si noti che non tutte le relazioni sono dimostrabili essere bidirezionali.
La figura e la sua didascalia sono riproduzioni della pubblicazione [3].

Come esempio di una ricerca, dove sono stati suggeriti criteri statistici concreti per l'unificazione in clan di gruppi di famiglie imparentate, può essere riportata la pubblicazione [4]. Tuttavia, un comune punto di vista per tale questione non esiste.

9. Altri metodi di analisi delle famiglie di proteine omologhe

In qualità di elemento costitutivo della ricerca nel processo di analisi delle famiglie di proteine omologhe, è possibile estrarre pattern conservativi di residui aminoacidici caratteristici per la famiglia in toto oppure per le sue distinte sottofamiglie. La presenza di tale pattern (consensus) nella sequenza aminoacidica della proteina può essere la base per l'accertamento della sua possibile appartenenza alla corrispondente famiglia (o sottofamiglia). Allo stesso modo possono presentare interesse posizioni caratteristiche nelle sequenze aminoacidiche, che permettono di distinguere i rappresentanti di diverse sottofamiglie. La ricerca di posizioni conservative nell'allineamento multiplo in un gruppo scelto di sequenze può essere eseguito automaticamente per mezzo del programma BioEdit. L'impiego di tali criteri per l'attribuzione della proteina alla già nota sottofamiglia può essere fondato se si ha un frammento relativamente corto di sequenza aminoacidica che non permette la conduzione di una analisi filogenetica.

Nell'analisi della famiglia di una proteina può inoltre rientrare la predizione ed i successivi confronti tra le strutture secondarie e terziarie dei suoi membri, sia tra di loro sia tra i rappresentanti delle famiglie imparentate.

Conclusioni

I risultati di un'analisi filogenetica possono essere pubblicati in un articolo a se stante dedicato all'evoluzione di una definita famiglia proteica. Come esempio di

tali lavori possono servire gli articoli [1] e [2]. In altri casi l'analisi filogenetica si presenta solo come elemento di una più ampia ricerca. Essa può essere condotta nella fase iniziale del lavoro che precede la messa in opera del problema sperimentale. Questo permetterà di scegliere adeguatamente un concreto rappresentante della famiglia della proteina che interessa per uno studio più dettagliato:

- predire la sua struttura tridimensionale e la sua organizzazione del dominio,
- predire la costruzione di un nucleo attivo e tracciare l'obiettivo di una mutagenesi sito-direzionale,
- predire le possibili attività enzimatiche.

L'analisi filogenetica può anche portare alla tappa conclusiva della ricerca, permettendo di determinare la posizione della proteina trovata e analizzata nel sistema gerarchico delle proteine già conosciute. In questo caso, l'albero filogenetico, che mostra la posizione della proteina analizzata può diventare un'ottima illustrazione per un articolo o una tesi. Si può comparare l'albero filogenetico di ogni dominio con l'albero evolutivo degli organismi-proprietari, ciò permetterà di trarre le conclusioni sulla natura evolutiva dei domini: quale ruolo hanno avuto la duplicazione, la perdita e la fusione dei geni, nonché i loro trasferimenti orizzontali.

Questo lavoro è stato finanziato con una borsa di studio del Presidente della Fed. Russa per i giovani ricercatori russi (MK-1461.2005.4) ed una borsa di studio della Fed. Russa (06-04-49079-a).

Letteratura

1. Naumoff D.G. 2004. Analisi filogenetica della famiglia GH27 α-galactosidase. Biologia molecolare. T.38. N.3. Pag.463-476. Abstract; PDF

2. Naumoff DG. 2005. GH97 is a new family of glycoside hydrolases, which is related tothe α-galactosidase superfamily. BMC Genomics.V.6. Art.112. Abstract; PDF

3. Rigden DJ. 2002. Iterative database searches demonstrate that glycoside hydrolase families 27, 31, 36 and 66 share a common evolutionary origin with family 13. FEBS
Lett. V.523. N.1-3. P.17-22. Abstract; PDF

4. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL and Bateman A. 2006.
Pfam: clans, web tools and services. Nucleic Acids Research. V.34. Database issue. D247-D251. Abstract; PDF