社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－BIO－4 (9)
2006／2／10

真核生物におけるドメインの組み合わせの系統解析

伊藤 真純 † 五斗 進 † 金久 實 †

† 京都大学化学研究所バイオインフォマティクスセンター

タンパク質のドメインは、タンパク質を構成する要素であり、タンパク質の機能、構造、進化の基本的な単位である。多くのタンパク質は複数のドメインから成り、このドメインの組み合わせにより機能の多様性が実現されている。これらマルチドメインタンパク質の形成はタンパク質機能の進化に大きな関わりがあると考えられている。これまで、ドメインの組み合わせは kingdom specific に起こることが指摘され、またその組み合わせと種特異的な機能との関連性について解析が成されてきた。今回、我々はより複雑なドメイン構成のタンパク質を持つ真核生物に関して、ドメインとその組み合わせの出現時期を推定し、真核生物におけるタンパク質ドメインの組み合わせの進化過程について調べた。

Phylogenetic Analysis of Domain Combinations in Eukaryotic Genomes

Masumi ITOH †, Susumu GOTO †, Minoru KANEHISA †,

† Bioinformatics Center, Institute for Chemical Research, Kyoto University

Protein domains are basic building blocks and can be units of the function, structure and evolution. A protein often consists of more than a domains and their combinations provide a broad functional spectrum and effective evolution of proteins. Previous work suggests that kingdom specific domain combinations, and species- or phylogenetic group- specific domain combinations cause particular biological mechanisms. Many multi-domain proteins are involved in regulatory systems of eukaryotes. Here, we have estimated specific domain combinations for various evolutionary related groups and obtained an outline of eukaryotic evolution of domain combinations.

1 はじめに

タンパク質のドメインは、タンパク質を構成する要素であり、タンパク質の機能、構造、進化の基本的な単位である。多くのタンパク質は複数のドメインから成り、ドメインの組み合わせは機能の多様性に大きく寄与しているとされている。

タンパク質を構成するドメインの組み合わせは真核生物、真正細菌、古細菌で kingdom specific に起こることが観察されており [1]、その組み合わせと種特異的な機能との関連性について解析されてきた。また、タンパク質ドメイン一つあたりの組み合わせのパートナーの数はべき乗則に従うことが報告されており [2, 3]、ドメイン自体の進化のメカニズムとあわせて、組み合わせの進化のメカニズムについても研究が行われている。マルチドメインタンパク質は、遺伝子の融合とドメインの欠如、内部での重複によって形成され、そのイベントが起こった過程を推定することでタンパク質の進化を解析することも試みられている [4]。また、タンパク質のドメインの出現時期と、イントロン-エキソン境界のコドンの読み枠やタンパク質上での配置に相関があることも分かっており [5]、ドメインの出現時期はドメインの機能・性質において非常に重要であると考えられる。

真核生物、とくに複雑な体制をもつ高等多細胞生物のタンパク質はより複雑なドメイン構成を持ち、多くのドメインが様々なタンパク質に繰り返し現れていることが近年のゲノムプロジェクトの結果解明されてきた [6]。ゲノムプロジェクトにより生物の持つ全遺伝子が明らかになることから、系統プロファイルと呼ばれるオーソログ遺伝子の有無を指標とした、生物の系統関係に基づくタンパク質の機能の解析が行われるようになってきた。真核生物でも、ゲノムプロジェクトの進行に伴い、ゲノムの決まった真核生物のオーソログの有無に基づいて、祖先生物でのオーソログの有無が推定され、遺伝子の獲得と欠損の過程について解析され

ている [7, 8, 9]。特に、この数年で、ヒトなどの哺乳類を含む多数のゲノム配列が決定されゲノムの決定された真核生物は急増しており現段階では 50 を超える真核生物のゲノム配列が決定されている。さらに、100 を超える真核生物のゲノムプロジェクトが進行中であり、高等真核生物の比較ゲノム解析はますます重要となると考えられる。本研究では、原核生物とくらべて比較的、系統関係が明らかである真核生物のドメインの組み合わせの進化と機能進化との関連性の解明を目指し、真核生物特有のタンパク質ドメインとその組み合わせの出現時期の推定を行った。この際、出現位置の推定には最節約法を用いた。これらの結果、各分岐点までに出現したドメインとそれらの組み合わせが推定され、ドメインの組み合わせの増大する過程の解析が可能になった。

2 方法

2.1 ドメインデータと系統関係

今回の解析では、KEGG GENES データベース、KEGG DGENES データベース [10] および Ensembl データベース [11] で公開されている 52 の真核生物のタンパク質配列を用いた (図 1)。これらの真核生物はゲノム配列もしくはドラフトゲノム配列の発表された種で、タンパク質配列が網羅的に予測され登録されている。KEGG GENES に登録されているタンパク質配列は、Pfam データベース [12] に登録されているドメイン領域が、HMMER[13, 14] を用いてアサインされており、KEGG DGENES および Ensembl から得られたタンパク質配列に関しては KEGG GENES と同じバージョンを用いて (Pfam ver.14) でドメインを定義した。今回、閾値として E-value $< 10^{-3}$ を用い、複数のドメインが重なって同じ領域に割り当てられたときは E-value の小さい方を採用した。同一のタンパク質内に二つの異なるドメインがアサインされた場合、その二つのドメインのペアを組み合わせとして定義した。このときペアとなるそれぞれのドメインが N 末側-C 末側のどちらにあるかを区別した。

今回は、ドラフトゲノムのみが決定された種が多く含まれていること、種の分岐時期が近くて分岐順が曖昧なことなどを考慮し、図 1 に示す多分木の系統樹を用いた。

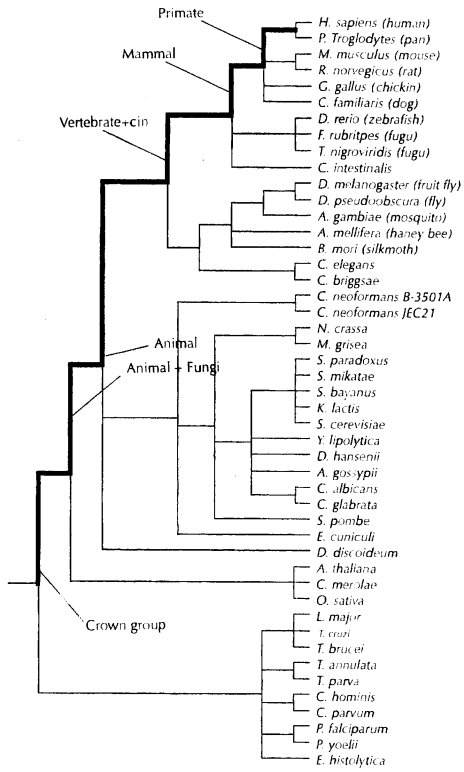

図 1: 今回使用した真核生物の系統関係
図中の太線はヒトにいたる系統

2.2 手法

2.2.1 最節約法による出現時期の推定

遺伝子が新しく出現することは進化の過程では 1 度しか起こらないと考えられるが、真核生物から原核生物、あるいは原核生物同士では水平伝播が比較的よく起こることが分かっているため、単純にある遺伝子をもつ全て生物の共通祖先ですでにその遺伝子が存在していたとは限らない。そのため Mirkin らは、既存生物種の系統関係とオーソログ遺伝子の分布から、進化の過程で起こる遺伝子の獲得 (出現・水平伝播)、欠損のシナリオを推測する手法を発表している [15]。この手法では、系統樹の各中間ノードに対応する祖先生物が、その親からオーソログ遺伝子を受け継いだ場合、受け継がなかった場合に場合分けし、またそれぞれの場合で、遺伝子の欠損・獲得が起こった場合のイベン

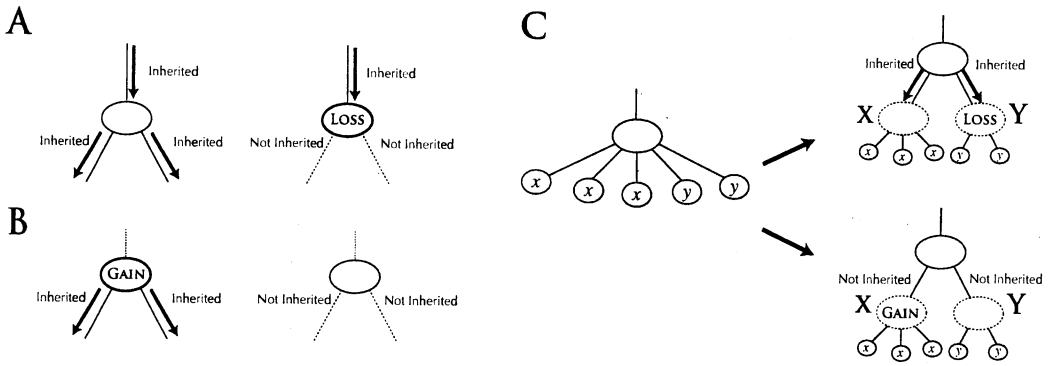

図 2: 最節約法による獲得・欠損のシナリオの推定

最節約法による遺伝子の出現時期を推定では、進化の過程での遺伝子の獲得と欠損のイベントが最少になるようなシナリオを推定することで行う。Mirkin らの手法 [15] では、系統樹内の中間ノードに対応する祖先生物について親ノードからオーソログ遺伝子を受け継いだ時 (A) と受け継がなかった時 (B) に場合分けし、それぞれの場合に関して、その中間ノード以下の遺伝子の獲得・欠損のイベント数を推定を推定している。親ノードから受け継いだ場合、遺伝子が子ノードに受け継がれる場合 (A 左) と、遺伝子の欠損が起こって受け継がれない場合 (A 右) でイベント数を計算し、より小さい方を採用する。親ノードから受け継がなかった場合も同様、獲得が起こって子ノードに受け継がれる場合と (B 左)、受け継がれない場合 (B 右) でイベント数の小さい方が採用される。この際、中間ノードで起こる獲得・欠損はイベント数に加えられる。また、獲得と欠損の頻度の比により加えられる値は重み付けされる。この過程はリーフとなる現存の生物から祖先生物へと再帰的に行われる。

今回、多分木に適用するためこの方法を改良した (C)。多分岐している中間ノードに関しては、その子ノードの系統関係が最もイベントが少ない形であると仮定し、子ノード (x、y) と中間ノードの間に遺伝子を子孫に受け継ぐノード X と受け継がないノード Y を考える。この際、親ノードから遺伝子を受け継いだ方がイベント数が小さくなる子ノード (x) を X の子孫、受け継がない方が小さくなるもの (y) を Y の子孫とする。対象の中間ノードから遺伝子が X、Y に受け継がれる場合 (C 右上)、Y で 1 度だけ欠損が起こるのが最小のイベント数となる。受け継がれない場合 (C 右下) も同様、X で 1 度だけ獲得が起こるのが最小である。また、X、Y 以下ではイベントが起こらないのが最もイベントの少ないシナリオであるので、系統関係にイベント数は依存しない。この後、X、Y を子ノードとして Mirkin らの手法を適用する。

ト数を計算し、イベントが最少になるシナリオを推定している (図 2A, B)。この際、遺伝子の獲得と欠損の起こる頻度の比が重要なパラメータとなる。今回、多分木にこの方法を適用するため、多分岐となる祖先生物から最もイベントの少なくなる分岐が起こったと仮定しする方法を用いた (図 2C)。この拡張により、進化関係が曖昧な場合にも適応できるようになる。

我々は、まず真核生物にのみ存在するドメインを推定するため、Mirkin らの手法に基づき 16S ribosomal RNA のアライメントから得られた真正細菌および古細菌の系統関係を基にそれぞれの共通祖先から存在するドメインを推定し、これら以外のドメインを真核生物特有のドメインとした。それに対して、真核生物間あるいは原核生物から真核生物への水平伝播はあまり起こらない [9] ことから、真核生物特有のドメイン出現時期はそのドメインを持つ全ての生物種の共通祖先であるとした。

ドメインの組み合わせは、遺伝子の融合、ドメインの挿入、欠如、配列内での重複によって形成される。前述のように真核生物において水平伝播による獲得は起こらないと考えられるが、同じ組み合わせのドメインの融合が別々の系統で独立して起こることは考えられる。そのため、我々はドメインの組み合わせの形成時期を上述の最節約法をもちいて推定した。

2.2.2 サブファミリーの定義

今回、タンパク質ドメインのサブファミリーの分岐時期の推定も行い、サブファミリーの出現時期と組み合わせの関連性についても解析を行った。真核生物特有のドメインについて、得られたドメインの配列を切り出し、HMMER の hmmalign を用いてそのドメインの HMM プロファイルに対し

表 1: 検出された真核生物のドメイン

	Pfam domain	protein
All	7,459	704,008
Assigned	4,315	403,329
Eukaryotic	3,104	274,289
Common	1,211	152,017

てアライメントを行い、マルチプルアラインメントを得た。HMM プロファイルに対してアライメントされずにギャップが挿入されている領域の配列を取り除き、QuickTree[16] を用いて UPGMA 法でタンパク質の階層的クラスタリングを行った。得られた樹形図の各分岐に関して、分岐以下に存在するタンパク質をもつ生物の共通祖先から、その分岐が種分岐による分岐か重複による分岐かを推定し、重複による分岐をサブファミリーの分岐として検出した。

3 結果

3.1 ドメインと組み合わせの出現時期の推定

今回、出現時期の推定の際、ドメインの獲得と欠損の頻度の比を 1〜3 として推定を行ったが、全体の傾向は変化しなかった。以下の結果は全て頻度の比を 3 とした場合であるが、本来この値はタンパク質ファミリーごとにも異なり、重要な問題であるためさらなる解析が必要である。原核生物のもつドメインの出現時期の推定の結果、真核生物のもつ 4,315 種のドメインのうち 1,211 種が真核生物と原核生物の共通祖先から存在していると推定され、残りの 3,104 種のドメインを真核生物特有のドメインとした（表 1)。

これらのドメインは 10,143 種類のドメインの組み合わせを形成しており、これらのドメインとその組み合わせの出現時期を推定した結果、573 種の組み合わせのみ、複数の祖先生物種で起こっていると推定され、多くの組み合わせは進化の過程で 1 度だけ起こっていると考えることが出来る。推測の結果を、ヒトに至る系での各祖先生物種についてまとめたものが図 3 である。各段階の祖先生物種がもっていたと考えられるドメインと組み合わせの総量と増加分を示している。図 3 には大まかな分岐年代を Million Years Ago[MYA] で示した。

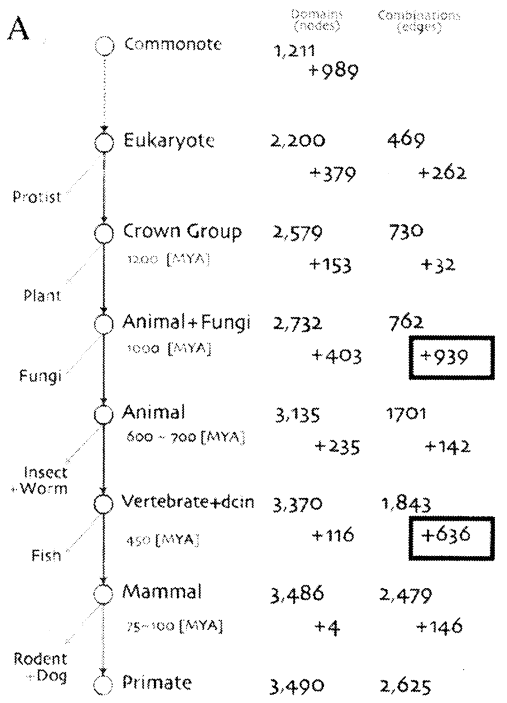

図 3: ドメインと組み合わせの増加

これらの結果から、真核生物のドメインは原核生物から真核生物の共通祖先が分岐後（Commonote〜Eukaryote）に大量に増えた以降は、動物と菌類の分岐後（Animal+Fungi〜Animal）でやや多めに増えているが、ほぼ一定の割合で緩やかに増加していると考えられる。一方、ドメインの組み合わせは動物と菌類の分岐後（Animal+Fungi〜Animal）に急激に増加しており、また哺乳類の分岐（Vertebrate+cin〜Mammal、cin はホヤを指す）の後にも急激に増加している。

3.2 出現時期と組み合わせの多様性

表 2 は図 3 の霊長類に存在するドメインの各出現時期ごとの内訳と、そのドメインがふくまれる組み合わせを示している。また Average は 1 つのドメインが含まれる平均的な組み合わせの数であるこの平均値は、ドメインをノード、ペアをエッジとしたネットワークの平均次数に相当する。ここにに示されるように、真核生物の共通祖先で出現したドメインと動物の共通祖先で出現したドメインは多くの組み合わせに含まれており、平均次数

表 2: 霊長類の持つドメインの出現時期

Emerge	Domain	Combination	Average
Commonote	1211	1567	1.29
Eukaryote	989	2285	2.31
Crown	379	268	0.71
Animal+Fungi	153	161	1.05
Animal	403	783	1.94
Vertebrate+cin	235	140	0.60
Mammal	116	46	0.40
Primate	4	0	0.00

表 3: 真核生物の共通祖先で出現したドメインのサブファミリーと組み合わせ

	Subfamily	Domain	Combination
Eukaryote	-	515	724
Crown	1,003	201	300
Animal+Fungi	1,011	172	397
Animal	1,769	261	762
Vertebrate+cin	2,483	194	424
Mammal	2,200	154	225
Primate	249	41	6

がそれぞれ 2.31、1.94 となっている。つまり、真核生物の共通祖先と動物の共通祖先で出現したドメインは他のドメインとペアを作りやすいと言うことを示している。

これらのタンパク質ドメインは Pfam データベースの定義に従いアサインしてきたが、これらのドメインは進化の過程で様々に分化しサブファミリーを形成する。上述の真核生物の共通祖先で出現したドメインもサブファミリーを多く形成している。表 3 に真核生物の共通祖先で出現したと考えられるドメインについて、階層的にクラスタリングの結果から推測された各祖先生物で出現したサブファミリーとそのサブファミリーを含むドメインの組合わせの数を示した。Subfamily の列は、各行の祖先生物で遺伝子重複により出現したと考えられるサブファミリー数を示しており、Domain は重複しサブファミリーを形成したと考えられる Pfam ドメインの種類を示している。この際、サブファミリー数は、同じ祖先生物で連続して複数回重複していると考えられているものに関しては、はじめの 1 回だけをカウントした。また、異なる祖先生物で複数回重複しているサブファミリーのドメインの組み合わせは、その組み合わせを持つサブファミリーのうち共通祖先で出現したサブファミリーでのみカウントした。表中の値は各行に対応する祖先生物のサブファミリーにユニークな組み合わせの数を示している。

4 考察

ドメインとその組み合わせの出現時期の推定の結果から、ドメインの増加に比較して動物と菌類の分岐以降、急激にドメインの組み合わせが増えていることが示された。これまで、体制が複雑な生物ほどドメイン構造が複雑になることは知られており、今回用いた動物は比較的高等な生物を多く含むため、その知見とよくあう。また、今回の結果でホヤおよびと脊椎動物の共通祖先からの哺乳類の分岐後に非常に多くのドメインの融合が見られた（図 3)。この増加はこの分岐後、魚類と哺乳類で共通に起こっていると考えられる。また、動物と菌類の分岐以降、ドメインの融合速度が加速しているとみることもできるかもしれない。

これらのドメインの組み合わせの多くは真核生物あるいは動物の共通祖先で出現したドメインを含み（表 2)、また真核生物の共通祖先で出現したドメインについては、動物分岐後に出現したサブファミリーがもっとも多様なドメインの組み合わせをもつこと（図 3）から、上述のドメインの組み合わせの増大に、これら動物分岐後に出現したファミリーあるいはサブファミリーが大きく寄与していることが強く示唆される。一方で、真核生物の祖先で出現したドメインは、脊椎動物とホヤの分岐後、あるいは哺乳類の分岐後には多数のサブファミリーが重複が起こっているが、これらにはユニークなドメイン構造を持つものが比較的少ない。また、重複しているドメインの種類も少なく、動物の分岐での様子とは異なり、ドメイン構造とそれぞれの分岐で起こった進化のメカニズムに対する知見が得られると期待される。

参考文献

[1] Apic, G., Gough, J. and Teichmann, S.: Domain combinations in archaeal, eubacterial and eukaryotic proteomes., *J Mol Biol*, Vol. 310, No. 2, pp. 311–25 (2001).

[2] Qian, J., Luscombe, N. and Gerstein, M.: Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model., *J Mol Biol*, Vol. 313, No. 4, pp. 673–81 (2001).

[3] Ye, Y. and Godzik, A.: Comparative analysis of protein domain organization., *Genome Res*, Vol. 14, No. 3, pp. 343–53 (2004).

[4] Bjorklund, S., Ekman, D., Light, S., Frey-Skott, J. and Elofsson, A.: Domain rearrangements in protein evolution., *J Mol Biol*, Vol. 353, No. 4, pp. 911–23 (2005).

[5] Vibranovski, S. and de Oliveira, d. S.: Signs of Ancient and Modern Exon-Shuffling Are Correlated to the Distribution of Ancient and Modern Domains Along Proteins., *J Mol Evol*, Vol. 61, No. 3, pp. 341–350 (2005).

[6] Koonin, E., Aravind, L. and Kondrashov, A.: The impact of comparative genomics on our understanding of evolution., *Cell*, Vol. 101, No. 6, pp. 573–6 (2000).

[7] Ogura, A., Ikeo, K. and Gojobori, T.: Estimation of ancestral gene set of bilaterian animals and its implication to dynamic change of gene content in bilaterian evolution., *Gene*, Vol. 345, No. 1, pp. 65–71 (2005).

[8] Makarova, K., Wolf, Y., Mekhedov, S., Mirkin, B. and Koonin, E.: Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell., *Nucleic Acids Res*, Vol. 33, No. 14, pp. 4626–38 (2005).

[9] Koonin, E., Fedorova, N., Jackson, J., Jacobs, A., Krylov, D., Makarova, K., Mazumder, R., Mekhedov, S., Nikolskaya, A., Rao, B., Rogozin, I., Smirnov, S., Sorokin, A., Sverdlov, A., Vasudevan, S., Wolf, Y., Yin, J. and Natale, D.: A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes., *Genome Biol*, Vol. 5, No. 2, p. R7 (2004).

[10] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG., *Nucleic Acids Res*, Vol. 34, No. Database issue, pp. D354–7 (2006).

[11] Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X., Flicek, P., Gr??f, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlic, A., Rae, M., Rios, D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C. and Hubbard, T.: Ensembl 2006., *Nucleic Acids Res*, Vol. 34, No. Database issue, pp. D556–61 (2006).

[12] Finn, R., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S., Sonnhammer, E. and Bateman, A.: Pfam: clans, web tools and services., *Nucleic Acids Res*, Vol. 34, No. Database issue, pp. D247–51 (2006).

[13] Eddy, S.: Hidden Markov models., *Curr Opin Struct Biol*, Vol. 6, No. 3, pp. 361–5 (1996).

[14] Eddy, S.: Profile hidden Markov models., *Bioinformatics*, Vol. 14, No. 9, pp. 755–63 (1998).

[15] Mirkin, B., Fenner, T., Galperin, M. and Koonin, E.: Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes., *BMC Evol Biol*, Vol. 3, p. 2 (2003).

[16] Howe, K., Bateman, A. and Durbin, R.: QuickTree: building huge Neighbour-Joining trees of protein sequences., *Bioinformatics*, Vol. 18, No. 11, pp. 1546–7 (2002).