FEBS Letters 583 (2009) 3589–3592

FEBS Letters

journal homepage: www.FEBSLetters.org

A kinome of 2600 in the ciliate *Paramecium tetraurelia*

Felix Bemm, Roland Schwarz [1], Frank Förster, Jörg Schultz *

Department of Bioinformatics, Biocenter, Am Hubland, 97074 Würzburg, Germany

ARTICLE INFO

Article history:
Received 3 August 2009
Revised 8 October 2009
Accepted 13 October 2009
Available online 17 October 2009

Edited by Takashi Gojobori

Keywords:
Kinase
Genome
Sequence analysis
Domain

ABSTRACT

Protein kinases play a crucial role in the regulation of cellular processes. Most eukaryotes reserve about 2.5% of their genes for protein kinases. We analysed the genome of the single-celled ciliate *Paramecium tetraurelia* and identified 2606 kinases, about 6.6% of its genes, representing the largest kinome to date. A gene tree combined with human kinases revealed a massive expansion of the calcium calmodulin regulated subfamily, underlining the importance of calcium in the physiology of *P. tetraurelia*. The kinases are embedded in only 40 domain architectures, contrasting 134 in human. This might indicate different mechanisms to achieve target specificity.

1. Introduction

Protein kinases are key players in cellular signal transduction. Together with their antagonists, the protein phosphatases, they tightly regulate the phosphorylation status, and thereby the activity, of their target proteins. Whereas the first years of protein kinase research focussed mainly on the experimental analysis of defined members of this protein family, the upcoming of completely sequenced genomes allowed for the 'in silico' characterisation of the complete kinase complement of an organism, the kinome. As a first in a growing list of eukaryotes, the yeast *Saccharmoyces cerevisiae* was sequenced. Of its roughly 6000 genes, 113 were protein kinases, totalling to about 2% [1]. With humans being so much more complex than yeast (at least from a humans perspective) one expected a substantially higher percentage of kinases in the human genome. First estimates were about 1000 kinases [2], later even went up to 2000 [3]. Therefore it was unexpected when, after sequencing the human genome, only 518 kinases were identified [4], roughly the same 2.5% as in yeast. Following analyses of further genomes like those of the sea urchin, *Drosophila melanogaster* and *Dictyostelium discoideum* revealed about the same percentage of kinases.

From there the analysis of the kinome of the ciliate *Tetrahymena thermophila* [5] gave a surprise. This single-celled organism harbours more than 1000 protein kinases, totalling up to 3.8% of the whole genome. This massive expansion was attributed mainly to an expansion of a single class of protein kinases and contrasted with the lack of classical protein tyrosine kinases. Here, we analysed the kinome of another ciliate, *Paramecium tetraurelia*, frequently used as a model organism for different aspects of cell biology [6,7]. Unlike *T. thermophila* this organism has undergone two additional rounds of whole genome duplications (WGDs) combined with phases of massive gene loss [8]. We address the question how this flexibility of gene content is reflected in the kinase complement.

2. Materials and methods

2.1. Identification

For the identification of protein kinase homologs, the kinomes of *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Homo sapiens* were downloaded from KinG [9] and those of *Monosiga brevicollis*, *Mus musculus*, *Strongylocentrotus purpuratus*, *Dictyostelium discoideum* and *Tetrahymena thermophila* from KinBase (http://www.kinase.com/kinbase/). Each sequence was blasted against the proteome of *P. tetraurelia* (downloaded from ParameciumDB [10], version 1.32) using standard parameters and an *E*-value cut-off of 10^{-3}. To identify the full length localisation of the kinase domain, each found protein was aligned against profile HMMs of serine/threonine and tyrosine kinases, respectively, downloaded from SMART [11] using the HMMer package [12].

2.2. Classification

To classify the protein kinase homologs, we used the multilevel HMM library Kinomer [13] with specific *E*-value cut-offs for each

* Corresponding author. Fax: +49 (0)931 318 4552.
E-mail address: Joerg.Schultz@biozentrum.uni-wuerzburg.de (J. Schultz).
[1] Present address: Cancer Research UK Cambridge Research Institute, Cambridge, UK.

0014-5793/$36.00
doi:10.1016/j.febslet.2009.10.029

kinase subfamily represented in the HMM library. Kinases which did not hit any HMM above the specific cut-offs were classified as "Others". Following, all classified kinases were blasted against all sequences used to build the Kinomer HMM library and the already classified kinases from KinBase. If a kinase hit a sequence that was classified differently, it was re-classified as "Others". The combination of those two methods resulted in a set of well-classified kinase homologs and a second set of homologs with no clear classification.

2.3. Transcriptome

To test for the presence of ESTs for the identified kinases, the coding sequence (CDS) of each kinase was blasted against all currently sequenced *P. tetraurelia* ESTs. An EST was assigned to a kinase, if the alignment contained a window of 100 base pairs containing 99 identical base pairs [14].

2.4. Domain analyses

The domain architecture of the kinases was predicted by Pfam [15]. Multiple consecutive occurrences of a domain were collapsed. All sequences with only a kinases domain were checked for novel domains. A domain was defined as a part of a protein homologous to at least one other protein with a different domain architecture. First, the kinase domain was deleted, leading N- and C-terminal fragments. These were dissected on coiled coil domains [16]. After filtering for compositional biased regions using seg, each resulting fragment >50 AA was blasted against a *P. tetraurelia* proteome without kinases. If a significant ($E < 10^{-3}$) hit against a known domain was found, the fragment was classified as divergent homolog. In the remaining cases the hit was classified as novel genetically mobile domain, if the homologous protein contained a known domain.

2.5. Pairwise distances

We calculated a matrix of pairwise distances from scores of pairwise alignments. Pairwise alignment scores (obtained from standard alignment with symmetric scoring matrices) are symmetric similarity scores, where the maximum possible alignment score of any given sequence is the score resulting from aligning the sequence with itself.

Matrices of pairwise alignment scores therefore have a tendency towards being diagonal dominant, which would lead to all eigenvalues being positive and leave the matrix as positive (semi-)definite (PSD, see, e.g. [17]). Following [18,19], it is possible to extract pairwise distances directly from PSD matrices, as they can be seen as matrices of pairwise dot products between objects in a possibly infinite-dimensional feature space.

Unfortunately matrices of pairwise alignment scores are not necessarily PSD. We therefore projected our alignment score matrices onto the next PSD according to the Frobenius norm $||A||_2 = (\sum_{ij} a_{ij}^2)^{1/2}$ [20]. We further normalized the feature space vectors implicitly via $K_0 = DKD$, where $D = diag(1/\sqrt{diag(K)})$ to remove length driven biases from the distance measure. The distance matrix was finally computed via $d(i,j)^2 = K'_{ii} - 2K'_{ij} + K'_{jj}$.

3. Results and discussion

3.1. Identification

Using a combination of Blast- and HMM-based annotations we identified 2606 unique protein kinases in the genome of *P. tetraurelia*. In relation to the 39 550 (July 2009) genes of *P. tetraurelia*, this amounts to roughly 6.6% of the whole gene content. Thus, the kinome of the single-celled protist *P. tetraurelia* outnumbers those of animals and fungi in both, total amount and fraction of genes. Even compared with *T. thermophila*, the closest sequenced sister taxon, containing 1069 protein kinases [5], the kinome is extended in absolute numbers and in percentage of the proteome. *P. tetraurelia* has undergone at least two rounds of WGD since its divergence from the last common ancestor of *T. thermophila* [8]. To test for the functionality of the kinases, we checked for the presence of catalytic residues [21]. We identified 587 kinases with substitutions in their catalytic sites (Table S1), 22% of the whole kinome. This percentage is roughly the same as in *T. thermophila*. Additionally, we compared the kinome to EST data. Sixteen percent of the kinases with substitutions were represented, contrasting 23% of the active kinases. Thus, some of the 'inactive' kinases might indeed be pseudogenes, but a substantial fraction should still be functional.

3.2. Classification

Already 20 years ago a classification of protein kinases was suggested [22] which was based on the sequences of roughly 100 kinases known at that time. Having an organism with 2606 kinases, disproportional extension of specific subfamilies could indicate which physiological processes regulated by kinases are of importance for the organism. We used a two-step approach for the classification by combining subfamily specific HMMs (Kinomer) [13] with reciprocal Blast searches against classified kinomes. The largest group we identified was the calcium calmodulin dependent subfamily with 970 members (Table 1). Indeed, calcium is involved in different cellular processes of *P. tetraurelia* like phagocytosis and membrane trafficking [23]. Already in the beginning of the 80s the localisation of calmodulin within the cilia was revealed [24]. Nearly 30 years later, with the genome at hand, the expansion of a kinase subfamily can be seen as an example of genome content reflecting the life history of an organism [25]. Further large subfamilies were the AGC (cyclic nucleotide-dependent families of PKG and PKA, β-adrenergic receptor kinase family, PKC family and ribosomal S6 kinase family) and the CMGC kinases (cyclin-dependent, mitogen-activated, glycogen-synthase and CDK-like kinases) with 635 and 322 members, respectively (Table 1). Contrasting these expansions, only one tyrosine kinase was identified (GSPATP00033466001). 380 Kinases were classified as "Others". This group represents kinases that were classified with profiles from subgroups already described as "Others", as well as kinases with no classification. As many kinases are receptors in metazoans, we checked for the presence of predicted transmembrane regions in the full length kinase sequences [26]. We identified 97 putative transmembrane kinases, 76 more than in *T. thermophila*. Still, this is a small fraction of the total number of kinases but consistent with a function of kinase receptors in cell–cell communication, not urgently needed in a single-celled organism.

3.3. Domain architectures

The regulation of protein kinase activity is frequently fine-tuned by additional protein domains. These domains can be responsible for the localisation, the interaction with other proteins or the direct activation, among others. Thus, the domain architectures (the linear order of domains in a protein) of a kinome could give a hint on how they are regulated and how they are integrated into the cellular network. The 2606 kinases of *P. tetraurelia* are composed by only 40 architectures (Table S2). This contrasts the human kinome with 134 different architectures, with splice variants counted as different architectures. Only 14 of the *P. tetraurelia* architectures can be found in human, 18 are *Paramecium* specific

Table 1
Classification of *P. tetraurelia* protein kinases.

Group	Kinomer	Kinomer and re-Blast	Substitution in active sites	Transcriptome		EST coverage in%		
				Active	Inactive	Overall (%)	Active (%)	Inactive (%)
AGC	635	442	62	107	12	19	19	19
RGC	1	0	1	0	0	0	0	0
TK	2	1	2	0	0	0	0	0
TKL	21	16	1	3	0	14	15	0
CAMK	970	863	93	196	14	22	22	15
STE	118	80	22	17	9	22	18	41
Other	380	9	328	5	51	15	10	16
CK1	165	164	31	17	3	12	13	10
CMGC	322	281	47	122	4	39	44	9
Sum	2615	1856	587	467	93	21	23	16

(Table S3). This indicates a large number of domain shuffling events in *Paramecium*. Using a small domain detection pipeline (see Section 2) one novel, *Paramecium* specific domain was identified in four proteins (GSPATP00028174001/529-588, GSPATP00037735001/9-102, GSPATP00036491001/4-98, GSPATP00011777001/1278-1372).

3.4. Phylogenetic tree

To compare the kinome of *P. tetraurelia* to that of human, we reconstructed a phylogenetic tree containing both. In a first step, a multiple alignment of these more than 3000 sequences was calculated. Because of the high divergence within the kinase family, this alignment did not reveal sufficient phylogenetic information. Accordingly, GBlocks [27] deleted all positions even at the lowest sensitivity. We therefore used pairwise distances (see Section 2) as the basis for the tree reconstruction with FastME [28]. The resulting tree reflected the previous subfamily classification, indicating that the pairwise distances carried sufficient phylogenetic information (Fig. 1). As expected, independent duplications happened in the line to humans and *P. tetraurelia*. Nine kinases contained two kinase domains. Whereas in one case the two

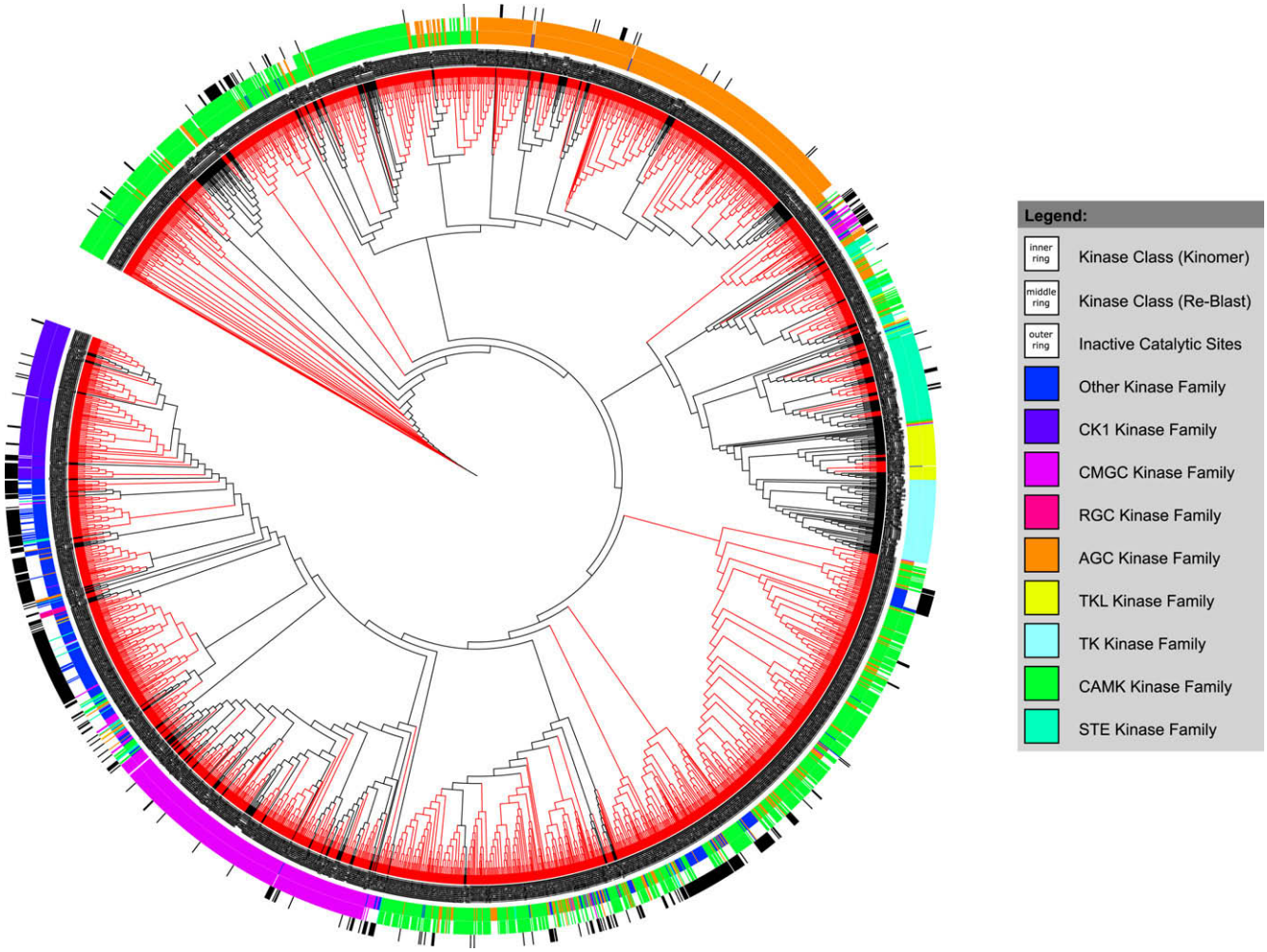

Fig. 1. Phylogenetic tree of *Paramecium tetraurelia* (red) and human (black) kinases – the tree was visualised using iTOL [30]. The classification of kinomer (inner coloured ring) and the combined classification with a reciprocal blast search (middle coloured ring) were mapped on the tree. The outer ring indicates kinases with substitutions at catalytic sites.

domains are highly similar, indicating a duplication event (GSPATP00015396A01), the others can be found in different positions in the phylogenetic tree (GSPATP00016887001, GSPATP00010360001, GSPATP00032460001, GSPATP00037688001, GSPATP00033260001, GSPATP00033979001, GSPATP00017902001, GSPATP0001946A001). These apparently arose via fusion events between different kinases.

Highlighting sequences with substitutions of catalytic sites revealed clusters enriched with probably inactive kinases, which, according to the classification, belong to different subfamilies or could not be classified. If a protein loses its function, it will evolve substantially faster than other proteins. In a tree, these fast evolving sequences can be drawn together, an artefact known as long-branch attraction. Thus, these clusters might have 'collected' kinases which are on their way to become a pseudogene. On the other hand, this indicated, that the other kinases with substitutions in catalytic sites might still perform non-catalytic, regulatory functions [21] or could have residual or scaffolding function.

4. Conclusions

With about 6.6% of the genome, *P. tetraurelia* contains the largest kinome to date. Based on the whole genome duplications, this ciliate had a straightforward mechanism to expand the absolute size of protein families. These duplications have been followed by massive loss of genes [5]. If there would be no preference to retain specific proteins, the relative size of protein families should stay constant. In contrast, we found an increase of the percentage of protein kinases compared to the close relative *T. thermophila*, which itself already has an expanded kinome. This increase could either indicate further, WGD independent duplications or a higher tendency to keep duplicated kinases. In any case, there seems to be an evolutionary pressure to sustain this vastly expanded kinome. Could this reflect a different mode of action of kinases in ciliates compared to metazoans? It is important that each kinase is directed exactly to its target protein. This could be assured by two approaches. Either the catalytic domains themselves have a high specificity or they are less specific but assisted by further domains and regulatory proteins. In the second case, especially when combined with differential splicing, a much smaller kinome might be sufficient. Indeed we found, compared to human, considerably fewer domain architectures. Furthermore, alternative splicing happens only rarely in *P. tetraurelia* [29]. Thus, *P. tetraurelia* might have tuned each kinase to its specific target. It would be interesting to test this hypothesis by the experimental characterisation of the phosphoproteome of *P. tetraurelia*.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2009.10.029.

References

[1] Hunter, T. and Plowman, G.D. (1997) The protein kinases of budding yeast: six score and more. Trends Biochem. Sci. 22, 18–22.
[2] Hunter, T. (1987) A thousand and one protein kinases. Cell 50, 823–829.
[3] Hunter, T. (1994) 1001 protein kinases redux—towards 2000. Semin. Cell Biol. 5, 367–376.
[4] Manning, G., Whyte, D.B., Martinez, R., Hunter, T. and Sudarsanam, S. (2002) The protein kinase complement of the human genome. Science 298, 1912–1934.
[5] Eisen, J.A., Coyne, R.S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J.R., Badger, J.H., Ren, Q., Amedeo, P., Jones, K.M., Tallon, L.J., Delcher, A.L., Salzberg, S.L., Silva, R.E., Haas, B.J., Majoros, W.H., Farzad, M., Carlton, J.M., Smith, R.K., Garg, J., Pearlman, R.E., Karrer, K.M., Sun, L., Manning, G., Elde, N.C., Turkewitz, A.P., Asai, D.J., Wilkes, D.E., Wang, Y., Cai, H., Collins, K., Stewart, B.A., Lee, S.R., Wilamowska, K., Weinberg, Z., Ruzzo, W.L., Wloga, D., Gaertig, J., Frankel, J., Tsao, C.-C., Gorovsky, M.A., Keeling, P.J., Waller, R.F., Patron, N.J., Cherry, J.M., Stover, N.A., Krieger, C.J., del Toro, C., Ryder, H.F., Williamson, S.C., Barbeau, R.A., Hamilton, E.P. and Orias, E. (2006) Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 4, e286.
[6] Plattner, H. (2002) My favorite cell—*Paramecium*. Bioessays 24, 649–658.
[7] Plattner, H. and Kissmehl, R. (2003) Molecular aspects of membrane trafficking in paramecium. Int. Rev. Cytol. 232, 185–216.
[8] Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Câmara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.-M., Kissmehl, R., Klotz, C., Koll, F., Le Mouël, A., Lepère, G., Malinsky, S., Nowacki, M., Nowak, J.K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Bétermier, M., Weissenbach, J., Scarpelli, C., Schächter, V., Sperling, L., Meyer, E., Cohen, J. and Wincker, P. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444, 171–178.
[9] Krupa, A., Abhinandan, K.R. and Srinivasan, N. (2004) KinG: a database of protein kinases in genomes. Nucleic Acids Res. 32, D153–D155.
[10] Arnaiz, O., Cain, S., Cohen, J. and Sperling, L. (2007) ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. Nucleic Acids Res. 35, D439–D444.
[11] Letunic, I., Doerks, T. and Bork, P. (2009) SMART 6: recent updates and new developments. Nucleic Acids Res. 37, D229–D232.
[12] Durbin, R. (2007) Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids, Cambridge Univ. Press, Cambridge.
[13] Martin, D.M.A., Miranda-Saavedra, D. and Barton, G.J. (2009) Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. Nucleic Acids Res. 37, D244–D250.
[14] Schultz, J., Doerks, T., Ponting, C.P., Copley, R.R. and Bork, P. (2000) More than 1,000 putative new human signalling proteins revealed by EST data mining. Nat. Genet. 25, 201–204.
[15] Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L.L. and Bateman, A. (2008) The Pfam protein families database. Nucleic Acids Res. 36, D281–D288.
[16] Lupas, A., van Dyke, M. and Stock, J. (1991) Predicting coiled coils from protein sequences. Science 252, 1162–1164.
[17] Golub, G.H. and VanLoan, C.F. (1996) Matrix Computations, Johns Hopkins Univ. Press, Baltimore.
[18] Shawe-Taylor, J. and Cristianini, N. (2006) Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge.
[19] Schölkopf, B. and Smola, A.J. (2002) Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA.
[20] Higham, N.J. (1988) Computing a nearest symmetric positive semidefinite matrix. Linear Algebra Appl. 103, 103–118.
[21] Pils, B. and Schultz, J. (2004) Inactive enzyme-homologues find new function in regulatory processes. J. Mol. Biol. 340, 399–404.
[22] Hanks, S.K., Quinn, A.M. and Hunter, T. (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science 241, 42–52.
[23] Fok, A.K., Aihara, M.S., Ishida, M. and Allen, R.D. (2008) Calmodulin localization and its effects on endocytic and phagocytic membrane trafficking in *Paramecium multimicronucleatum*. J. Eukaryot. Microbiol. 55, 481–491.
[24] Maihle, N.J., Dedman, J.R., Means, A.R., Chafouleas, J.G. and Satir, B.H. (1981) Presence and indirect immunofluorescent localization of calmodulin in *Paramecium tetraurelia*. J. Cell Biol. 89, 695–699.
[25] Emes, R.D., Goodstadt, L., Winter, E.E. and Ponting, C.P. (2003) Comparison of the genomes of human and mouse lays the foundation of genome zoology. Hum. Mol. Genet. 12, 701–709.
[26] Sonnhammer, E.L., Heijne, G. von and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology, Vol. 6, pp. 175–182.
[27] Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.
[28] Desper, R. and Gascuel, O. (2004) Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol. Biol. Evol. 21, 587–598.
[29] Jaillon, O., Bouhouche, K., Gout, J.-F., Aury, J.-M., Noel, B., Saudemont, B., Nowacki, M., Serrano, V., Porcel, B.M., Ségurens, B., Le Mouël, A., Lepère, G., Schächter, V., Bétermier, M., Cohen, J., Wincker, P., Sperling, L., Duret, L. and Meyer, E. (2008) Translational control of intron splicing in eukaryotes. Nature 451, 359–362.
[30] Letunic, I. and Bork, P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127–128.