Cellulosome-like sequences in *Archaeoglobus fulgidus*: an enigmatic vestige of cohesin and dockerin domains

Edward A. Bayer[a], Pedro M. Coutinho[b,1], Bernard Henrissat[b,*]

[a] *Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot 76100, Israel*
[b] *Architecture et Fonction des Macromolécules Biologiques - CNRS-IFR1, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France*

Received 21 September 1999; received in revised form 17 November 1999

Edited by Matti Saraste

Abstract The distribution of cellulosomal cohesin domains among the sequences currently compiled in various sequence databases was investigated. Two cohesin domains were detected in two consecutive open reading frames (ORFs) of the recently sequenced genome of the archaeon *Archaeoglobus fulgidus*. Otherwise, no cohesin-like sequence could be detected in organisms other than those of the Eubacteria. One of the *A. fulgidus* cohesin-containing ORFs also harbored a dockerin domain, but the additional modular portions of both genes are undefined, both with respect to sequence homology and function. It is currently unclear what function(s) the putative cohesin and dockerin-containing proteins play in the life cycle of this organism. In particular, since *A. fulgidus* contains no known glycosyl hydrolase gene, the presence of a cellulosome can be excluded. The results suggest that cohesin and dockerin signature sequences cannot be used alone for the definitive identification of cellulosomes in genomes.

Key words: Cellulosome; Cohesin domain; Dockerin domain; Archeon; Modular protein

1. Introduction

Cellulosomes are multi-enzyme complexes produced in various cellulolytic microorganisms [1], designed to hydrolyze efficiently plant cell wall polysaccharides. The best characterized cellulosome systems are those from the Clostridiae. In these, a central 'scaffoldin' subunit integrates catalytic subunits into the complex by virtue of two complementary types of domain, the *cohesins* on scaffoldin and the *dockerins* on the catalytic subunits [2,3]. The calcium-mediated cohesin-dockerin interaction appears to dictate cellulosome formation and cohesin and dockerin-like stretches in newly sequenced proteins have been considered to indicate the presence of cellulosomes.

In the present communication, we report the presence of a single dockerin and two cohesin-like sequences in the recently sequenced genome of the archaeon *Archaeoglobus fulgidus* [4]. The detected domains do not appear to be associated with known cellulosome-related components (e.g. scaffoldins, glycosyl hydrolases or anchoring proteins). Moreover, one of the parent coding regions contains a dockerin domain together with one of the cohesin domains.

2. Materials and methods

2.1. Sequences

The GenBank accession codes used for sequence analysis are as follows: scaffoldin proteins from *Clostridium thermocellum* (Clotm_CipA, L08665), *Clostridium cellulolyticum* (Cloce_CipC, U40345) and *Clostridium cellulovorans* (Clocl_CbpA, M73817); *Clostridium thermocellum* anchoring proteins SdbA, Orf2p and OlpB (U49980 and X67506); *C. thermocellum* outer-layer protein component A (OlpA, X67506); and *C. cellulolyticum* cohesin-containing protein gene (ORFX, AF081458). GenBank accession codes for the enzymes *C. thermocellum* CelS and *C. cellulolyticum* CelA are L06942 and M93096, respectively.

2.2. Sequence analysis methods

Gapped-BLAST and PSI-BLAST [5] searches for dockerin and cohesin modules in GenBank as well as unfinished genomes were performed through the NCBI server at URL http://www.ncbi.nlm.nih.gov/BLAST. Multiple sequence alignments were performed using ClustalW [6] and refined manually against the known three-dimensional (3D) structure of the cohesin 2 of *C. thermocellum* scaffoldin [7]. Secondary structure predictions were carried out using the PredictProtein server at URL http://dodo.cpmc.columbia.edu/predictprotein [8]. Phylogenetic analysis and construction of the phylogenetic tree was obtained by alignment of the different cohesin segment sequences with ClustalW [6] followed by tree reconstruction by PUZZLE [9], using quartet sampling and neighbor joining options. The Blosum62 matrix was used in both cases. In order to identify related sequences, all the putative coding regions of *A. fulgidus* have been BLASTed against a library of over 2500 glycoside hydrolases compiled in the laboratory and covering the 77 families of glycoside hydrolases presently known [10]. The significance of the alignments between *A. fulgidus* putative cohesins and dockerin with cellulosomal cohesins and dockerins was assessed using the PCOMPARE program of PC/Gene (Intelligenetics), applying Dayhoff's matrix and 100 randomizations.

3. Results

Two cohesin-like sequences were detected in two contiguous open reading frames (ORFs), 2375 and 2376, in the *A. fulgidus* genome (GenBank accession number AE001112). One of the putative proteins (ORF 2375) contains a cohesin module and a dockerin module in tandem (Fig. 1). The residual N-terminal portion of the protein (274 amino acid residues) exhibits no detectable similarity with any known gene and no signal peptide was apparent. The second putative protein (ORF 2376) contains a potential signal peptide (or transmembrane segment), a cohesin-like module and a short (~30 residues) C-terminal segment of unknown function that bears similarity with portions of several other ORFs (AF0275, AF0946, AF1949, AF1487) from the *A. fulgidus* genome. It is interest-

*Corresponding author. Fax: (33)-491-16 45 36.
E-mail: bernie@afmb.cnrs-mrs.fr

[1] Present address: Instituto Superior Técnico - Secção de Biotecnologia, Av. Rovisco Pais, 1049-001 Lisbon, Portugal.

PII: S0014-5793(99)01634-8

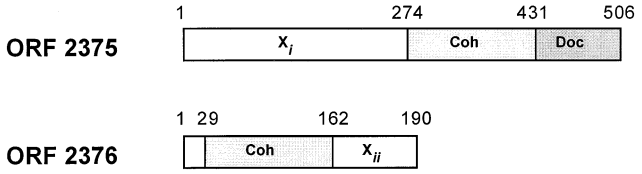

Fig. 1. Schematic representation of the modular structure of ORFs 2375 and 2376 from *A. fulgidus*. Both putative proteins contain cohesins (Coh). ORF 2375 also contains a dockerin (Doc) and a relatively large region (X_i) which failed to align against known proteins. ORF 2376 includes a short stretch of unknown function (X_{ii}) which aligns against similar stretches of four other predicted coding regions from the same genome.

ing to note that this short segment appears at the C-terminus in all these ORFs.

The similarity between the *A. fulgidus* cohesin and dockerin sequences and their clostridial counterparts was judged significant with scores ranging from 2.7 to 6.5 S.D.s above the average of 100 random runs. We have searched for other cohesin and dockerin-like sequences in GenBank, available at the Institute for Genomic Research (July 16, 1999), which currently includes 18 complete genomes and 35 unfinished genomes. Except for cohesin and dockerin sequences found

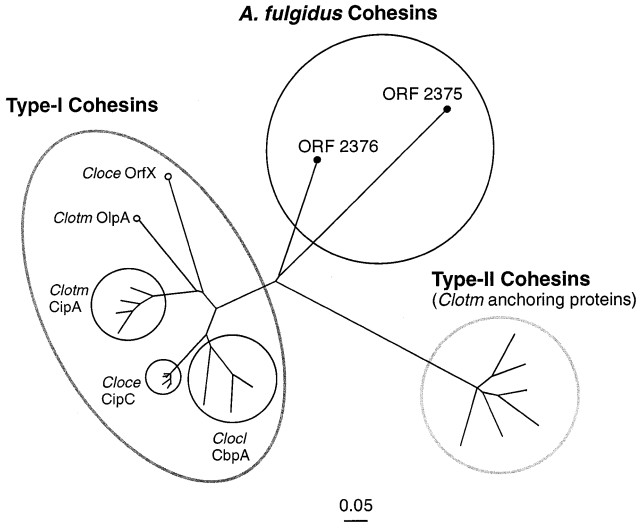

Fig. 2. Unrooted phylogenetic tree of the cohesins from *A. fulgidus* versus those from selected Clostridiae. The type-I cohesins are derived from the scaffoldins of *C. thermocellum* (*Clotm* CipA), *C. cellulovorans* (*Clocl* CbpA) and *C. cellulolyticum* (*Cloce* CipC) and from non-scaffoldin proteins (*Clotm* OlpA and *Cloce* OrfX). The type-II cohesins are derived from *C. thermocellum* anchoring proteins (SdbA, OlpB and Orf2p). The scale bar indicates the number of substitutions per position as given by PUZZLE.

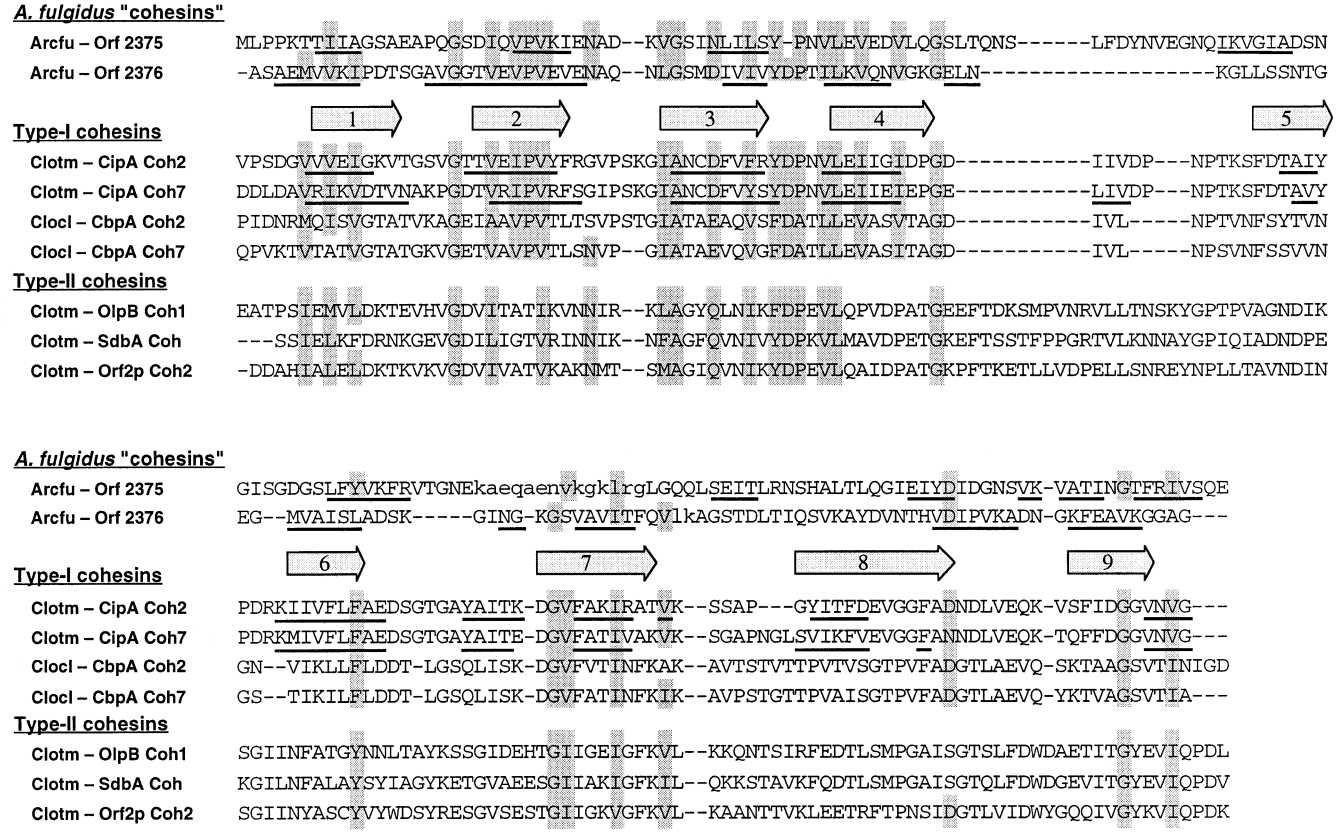

Fig. 3. Multiple sequence alignment of the *A. fulgidus* cohesins with selected clostridial type-I and type-II cohesins. Shown for comparison are type-I cohesin sequences from the cellulosomal scaffoldins of *C. thermocellum* and *C. cellulovorans*. Also shown are examples of the known type-II cohesins from *C. thermocellum* anchoring proteins SdbA, OlpB and Orf2p. The shaded residues indicate conservation (identity or group-specific similarity) at the given position of the *A. fulgidus* ORF(s) compared to the type-I and/or type-II cohesins. The numbered arrows indicate the positions of the β-strands from the known structures of cohesins 2 and 7 from the *C. thermocellum* scaffoldin. Underlined sequences of the *A. fulgidus* sequences show predicted β-strand regions and lower case lettering indicates predicted α-helical regions, according to the secondary structure prediction using the PredictProtein server. For comparison, a similar 'retrospective' view of the *C. thermocellum* cohesins 2 and 7 is provided: the positions of the predicted β-strands are underlined and should be compared with the actual, crystallographically determined positions (designated by the arrows).

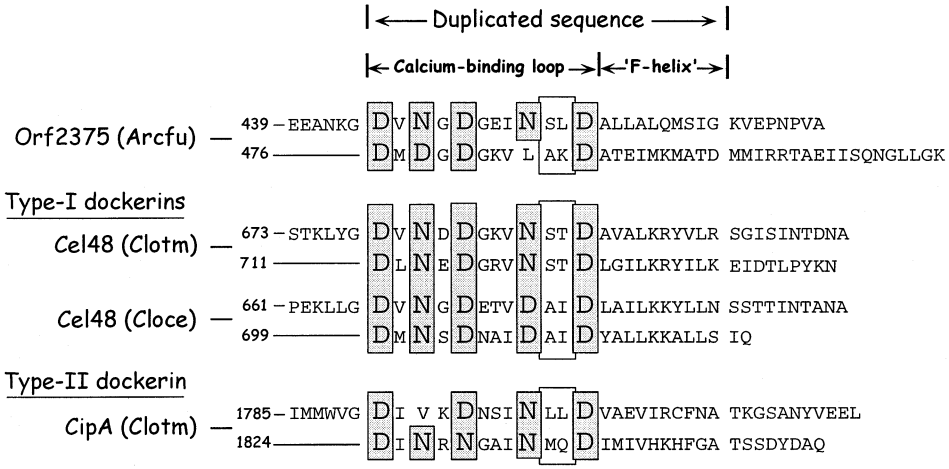

Fig. 4. Alignment of the dockerin from ORF 2375 of the *A. fulgidus* genome with typical type-I and type-II dockerins. Shown for comparison are the type-I dockerins of the family-48 cellulosomal enzymes from *C. thermocellum* and *C. cellulolyticum* and the type-II dockerin from the *C. thermocellum* scaffoldin (CipA). The positions of the repeated sequences, their calcium-binding loops and putative 'F-helices' are indicated. The proposed calcium-binding residues are emphasized in shaded boxes and predicted recognition residues are denoted in unshaded boxes.

in the putative cellulosome gene cluster of the *Clostridium acetobutylicum* genome, related sequences were not detected in any of the other available genomes.

Phylogenetic analysis indicates that the two cohesins of *A. fulgidus* occupy a separate branch of the cohesin tree (Fig. 2). The cohesins are located in an intermediate position between the type-I and type-II cohesins, but appear to have diverged considerably from each other.

A closer inspection of the sequences of the *A. fulgidus* cohesins versus those of the type-I and type-II cellulosomal cohesins is shown in Fig. 3. The *A. fulgidus* cohesins have been aligned in the figure against examples of both type-I and type-II cellulosomal cohesins, taking into account the secondary structure elements determined from the known crystal structures of cohesins 2 and 7 from the *C. thermocellum* scaffoldin subunit [7,11]. The sequence alignment of the *A. fulgidus* cohesin sequences is mirrored by the predicted secondary structural elements, which indicate a mainly-β or all-β architecture. The prediction is in close agreement with the predicted secondary structure of cellulosomal cohesins 2 and 7 of *C. thermocellum* as well as with the experimentally determined secondary structure derived from their 3D structures [7,11].

A typical bacterial dockerin includes an internal duplicated segment (hence registration as two copies in the PFAM server at URL http://pfam.wustl.edu). In fact, both copies of the duplicated sequence are required to form the functional dockerin unit, since binding to the cohesin is dependent on the presence of both repeats [12,13]. The *A. fulgidus* dockerin-like sequence displays standard features consistent with a dockerin domain. The region aligns closely with both type-I and type-II dockerins derived from known cellulosomal components (Fig. 4), and thus resembles the proposed F-helix variation [14] of the EF-hand motif of calcium-binding proteins, e.g. calmodulin and troponin C. Most of the putative calcium-binding residues are present in the anticipated position with one exception (L484). Such substitutions have been observed occasionally among dockerin sequences and most likely have a relatively minor effect on calcium-binding. As evident from the *A. fulgidus* ORF 2375 sequence, the residues of the two segments that align with the 'F-helix' cannot be considered a strict repeat. Nonetheless, secondary structure prediction indicates that both segments would be expected to form α-helices with the highest reliability index of prediction (data not shown). Finally, the putative recognition residues [14] appear to be distinct from those of the dockerins from the cellulosomal enzymes.

4. Discussion

Cellulosomal scaffoldins from cellulolytic microorganisms are composed of multiple repeats of cohesins and other functional modules. The cohesins bind strongly to the complementary dockerin domains of the cellulosomal enzymes, and the resultant type-I cohesin-dockerin interaction serves to integrate the enzymes into the complex [15]. A second type of cohesin-dockerin interaction has also been described for the *C. thermocellum* cellulosome. In this case, a type-II dockerin on scaffoldin binds tightly to type-II cohesins borne by a series of surface-bound anchoring proteins, thereby incorporating the cellulosome onto the cell surface in this bacterium [16]. Recent evidence indicates that the type-II cohesin-dockerin interaction may be typical for other cellulolytic species [17].

Thus far, cohesin and dockerin modules have been found only in anaerobic cellulolytic microorganisms. This study is the first instance of the detection of such sequences in an organism which is not cellulolytic. In this regard, BLAST/PSI-BLAST searches with all known glycoside hydrolase families failed to detect any significant match on the *A. fulgidus* genome. It is also noteworthy that although *A. fulgidus* is apparently devoid of glycosidases, its genome encodes at least 14 glycosyl transferases [18]. It is thus intriguing to consider that this organism apparently possesses all the necessary elements to make oligo and/or polysaccharides but would be unable to hydrolyze them.

The presence of cohesin and dockerin-like sequences in the *A. fulgidus* genome raises several basic questions, including how did they get there and what do they do?

Regarding their function, cellulosome-related cohesin-dockerin pairs simply bring together two different protein compo-

nents in a high-affinity protein-protein binding interaction. If we assume that the *A. fulgidus* dockerin will also selectively bind to a cohesin, then it may serve either to bind to the cohesin of ORF 2376 and/or to the cohesin of its own parent protein (ORF 2375). In the absence of defined functions for these putative proteins, it is difficult to gauge the logic of self-binding by ORF 2375. The only other known example of cohesin and dockerin domains on the same protein is the *C. thermocellum* scaffoldin, wherein the binding specificity differs, thus allowing multiple integration of enzymes and selective attachment of the cellulosome to the cell surface. If this is the case for ORF 2375, it is still difficult to reconcile the necessity for a single cohesin-dockerin pair between ORF 2375 and 2376. Further speculation should await biochemical evidence concerning the cohesin-dockerin binding specificity (if any) and functions of the other modules of these two ORFs.

Regarding their origin, it has long been suspected that the modules of cellulosomes and glycoside hydrolases can be transmitted by horizontal gene transfer among phylogenetically unrelated microorganisms that share the same ecosystem. It can thus be argued that a cellulolytic associate of *A. fulgidus* may have provided the necessary genetic material for incorporation of the cohesin and dockerin sequences into the genome. In this context, it is interesting to note that although *A. fulgidus* is taxonomically distant from the Clostridiaceae, its two cohesins are as similar to the clostridial type-I and type-II cohesins as the latter are to each other. Moreover, cohesin and dockerin-like sequences have not been detected in any other known archaeal genome, suggesting that such sequences may not be common to the Archaea.

The results suggest that cohesin and dockerin signature sequences cannot be used alone for definitive identification of cellulosomes in genomes. More generally, this implies further that prediction of the function of ORFs in genomes, simply on the basis of sequence homology, can often be erroneous, particularly in the case of modular proteins harboring domains that are subject to frequent horizontal exchange among organisms.

The current overall view of the universal phylogenetic tree of life, based on ribosomal RNA analysis, indicates three distinct primary groups: the Bacteria, the Archaea and the Eukarya [19,20]. The placement of the Archaea is particularly intriguing, since most of the information-processing components among its members resemble those of the Eukarya, whereas most of the metabolic components are similar to those of the Bacteria. In most schemes, the Archaea are usually considered to ally with the Eukarya in the prokaryote-to-eukaryote transition, although the identity of a universal ancestor and definitive phylogenetic tree of life is still subject to controversy [21,22]. Nevertheless, we hypothesize that the cohesin and dockerin-like sequences in the *A. fulgidus* genome may represent remnants of bacterial cellulosome genes acquired via horizontal transfer, rather than symbolizing primordial components of archaeal origin recruited by the cellulolytic bacteria.

Acknowledgements: This work was funded by contract BIO4-97-2303 of the European Commission. A TMR Grant (BIO4-97-5077) to P.M.C. is gratefully acknowledged.

References

[1] Bayer, E.A., Chanzy, H., Lamed, R. and Shoham, Y. (1998) Curr. Opin. Struct. Biol. 8, 548–557.
[2] Bayer, E.A., Morag, E. and Lamed, R. (1994) Trends Biotechnol. 12, 379–386.
[3] Bayer, E.A., Shimon, L.J., Shoham, Y. and Lamed, R. (1998) J. Struct. Biol. 124, 221–234.
[4] Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., Richardson, D.L., Kerlavage, A.R., Graham, D.E., Kyrpides, N.C., Fleischmann, R.D., Quackenbush, J., Lee, N.H., Sutton, G.G., Gill, S., Kirkness, E.F., Dougherty, B.A., McKenney, K., Adams, M.D., Loftus, B. and Venter, J.C. (1997) Nature 390, 364–370.
[5] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[6] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.
[7] Shimon, L.J., Bayer, E.A., Morag, E., Lamed, R., Yaron, S., Shoham, Y. and Frolow, F. (1997) Structure 5, 381–390.
[8] Rost, B. (1997) Methods Enzymol. 266, 525–539.
[9] Strimmer, K. and von Haeseler, A. (1996) Mol. Biol. Evol. 13, 964–969.
[10] Coutinho, P.M. and Henrissat, B. (1999) in: Recent Advances in Carbohydrate Bioengineering (Gilbert, H.J., Davies, G., Henrissat, B. and Svensson, B., Eds.), The Royal Society of Chemistry, Cambridge (in press).
[11] Tavares, G.A., Béguin, P. and Alzari, P.M. (1997) J. Mol. Biol. 273, 701–713.
[12] Fierobe, H.-P., Pagès, S., Belaich, A., Champ, S., Lexa, D. and Belaich, J.-P. (1999) Biochemistry 38, 12822–12832.
[13] Lytle, B. and Wu, J.H.D. (1998) J. Bacteriol. 180, 6581–6585.
[14] Pagès, S., Belaich, A., Belaich, J.-P., Morag, E., Lamed, R., Shoham, Y. and Bayer, E.A. (1997) Proteins 29, 517–527.
[15] Salamitou, S., Tokatlidis, K., Béguin, P. and Aubert, J.-P. (1992) FEBS Lett. 304, 89–92.
[16] Leibovitz, E. and Béguin, P. (1996) J. Bacteriol. 178, 3077–3084.
[17] Bayer, E.A., Ding, S.Y., Shoham, Y. and Lamed, R. (1999) in: Genetics, Biochemistry and Ecology of Cellulose Degradation (Ohmiya, K., Hayashi, K., Sakka, K., Kobayashi, Y., Karita, S. and Kimura, T., Eds.), pp. 428–436, Uni Publishers, Tokyo.
[18] Coutinho, P.M. and Henrissat, B. (1999) J. Mol. Microbiol. Biotechnol. (in press).
[19] Olsen, G.J. and Woese, C.R. (1997) Cell 89, 991–994.
[20] Brown, J.R. and Doolittle, W.F. (1997) Microbiol. Mol. Biol. Rev. 61, 456–502.
[21] Brinkmann, H. and Philippe, H. (1999) Mol. Biol. Evol. 16, 817–825.
[22] Doolittle, W.F. (1999) Science 284, 2124–2129.