Research

NEAT: a domain duplicated in genes near the components of a putative Fe^{3+} siderophore transporter from Gram-positive pathogenic bacteria

Miguel A Andrade*†, Francesca D Ciccarelli*†, Carolina Perez-Iratxeta*† and Peer Bork*†

Addresses: *European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany. †Department of Bioinformatics, Max Delbruck Center for Molecular Medicine, 13092, Berlin-Buch, Germany.

Correspondence: Miguel A Andrade. E-mail: andrade@embl-heidelberg.de

Published: 15 August 2002

Genome ***Biology*** 2002, **3(9)**:research0047.1–0047.5

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/9/research/0047

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 24 April 2002
Revised: 17 May 2002
Accepted: 5 June 2002

Abstract

Background: Iron uptake from the host is essential for bacteria that infect animals. To find potential targets for drugs active against pathogenic bacteria, we have searched all completely sequenced genomes of pathogenic bacteria for genes relevant for iron transport.

Results: We identified a protein domain that appears in variable copy number in bacterial genes that are usually in the vicinity of a putative Fe^{3+} siderophore transporter. Accordingly, we have denoted this domain NEAT for 'near transporter'. Most of the bacterial species containing this domain are pathogenic. Sequence features indicate that the domain is anchored to the extracellular side of the membrane. The domain seems to be under high selective pressure for rapid independent duplications that are typical of sequences involved in signaling and binding.

Conclusions: The NEAT domain might be functionally related to iron transport. The taxonomic specificity of this domain and its predicted extracellular position could make it an interesting target for designing new drugs against some highly pathogenic bacteria.

Background

Iron transport into the cell is very important for the growth of an organism. Pathogenic bacteria, which have to survive within an animal, are able to sequester iron from the iron-containing proteins of the host by secreting siderophores that have a higher affinity for the iron (reviewed in [1]). Then, a specific transport system imports the iron-siderophore complex back into the bacterial cytoplasm. The disruption of this uptake function in bacteria is likely to be a good strategy in fighting infectivity. We searched the genomic neighborhood of putative Fe^{3+} siderophore transporters in pathogenic bacteria in order to identify genes that could be associated with this functionality and thus constitute targets for therapy against disease. As a result of our analysis, we characterized a highly duplicated domain that we propose as a receptor for an iron complex.

Results and discussion

Survey for putative Fe^{3+} siderophore transporters in complete bacterial genomes

In order to find proteins related to iron transport in pathogenic bacteria, we first scanned complete genomes of pathogenic bacteria for sequences homologous to those encoding

the three currently known *Escherichia coli* Fe^{3+} siderophore transporters: the Fe^{3+} dicitrate transport complex [2], the Fe^{3+} enterobactin transport complex [3], and the Fe^{3+} hydroxamate transport complex [4]. These transporters import iron from the periplasm into the cytoplasm of *E. coli*, expending ATP. Several components of the putative transporter were found in contiguous genomic positions of four pathogenic Gram-positive bacteria, three of which are associated with food-borne diseases (*Listeria monocytogenes, Clostridium perfringens,* and *Staphylococcus aureus*). In humans, the fourth bacterium, *Staphylococcus pyogenes*, produces pharyngitis, impetigo, toxic shock syndrome, necrotizing fasciitis, rheumatic fever, and acute glomerulonephritis.

Gene neighborhood

In order to find genes associated with the putative Fe^{3+} siderophore transporter that could be characteristic of the pathogenic species, we analyzed the genomic neighborhood of the transporter in complete genomes. The repeated presence of neighboring gene pairs across different species permits us to reach conclusions about the possible functional association of the paired genes [5-7].

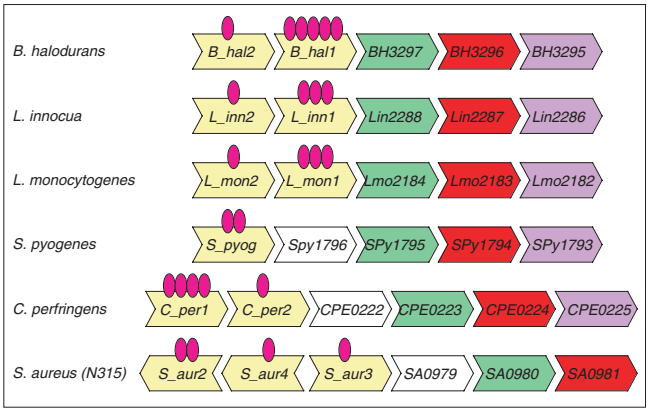

Figure 1
Conserved genome organization around the components of a putative Fe^{3+} siderophore transporter (from STRING [26,27] or from the literature when the genome is not present in STRING). Each gene is represented with an arrow-shaped box (as in STRING) pointing in the direction of transcription. The genes in yellow contain the NEAT domain; the pink ovals indicate the number of occurrences of the domain within the gene. Genes in green are related to *Escherichia coli fecB, fepB, or fhuD* (Fe^{3+} siderophore transporter, periplasmic component). Genes in red are related to *fecC, fepG, fhuB* (Fe^{3+} siderophore transporter, transmembrane component). Genes in violet are related to *fecE, fepC, fhuC* (Fe^{3+} siderophore transporter, ATPase component). Remaining genes (in white) did not show significant sequence similarity to any of the other genes displayed in the figure. Codes for genes containing the NEAT domain are shown in Table 1. For the neighboring genes the names used correspond to those from the corresponding genomic project. Spy1797 should be shown between S_pyog (Spy1798) and Spy1796, but is not displayed here because there was no such corresponding entry in GenBank. Note that there was no neighboring ATPase (violet gene) in *S. aureus*; the most similar sequence in this species is SA0602 (not shown).

The examination of the genomic neighborhood of the transporter indicated the correlated presence of genes containing a conserved domain, with a variable copy number (from one up to five) within the same sequence. We therefore denoted this newly described domain as NEAT, for 'near transporter'. A similar correlation between the transporter and genes containing the NEAT domain was also found in the nonpathogenic species *Listeria innocua* and *Bacillus halodurans* (see Figure 1).

The search for homologous sequences in the whole protein database added one more sequence from *S. aureus*, and a short protein corresponding to the middle part of the domain in the virulence plasmid pXO1 of *Bacillus anthracis*, which is essential for the manifestation of the disease anthrax. Both genes are apparently not physically close to transport-related genes.

The alignment of all identified instances of the NEAT domain (Figure 2) indicates a conserved region of about 125 amino acids. The predicted secondary structure (using PHD [8]) suggests that this domain is mostly composed of beta strands. The NEAT domain appears in combination with other domains and sequence features in various proteins (see Figure 3). A distinctive feature of most of these proteins is the prediction of an amino-terminal signal sequence and a

Table 1

Codes for genes containing the NEAT domain

Gene code	Protein accession number*	Bacterium	Reference
B_hal1	SP:Q9K7R1	*B. halodurans*	[20]
B_hal2	SP:Q9K7R0	*B. halodurans*	[20]
C_per1	GB:18143877	*C. perfringens*	[21]
C_per2	GB:18143878	*C. perfringens*	[21]
S_pyog	SP:Q99YA0	*S. pyogenes*	[22]
L_inn1	SP:Q92916	*L. innocua*	[23]
L_inn2	SP:Q92915	*L. innocua*	[23]
L_mon1	GB:16804224	*L. monocytogenes*	[23]
L_mon2	GB:16411656	*L. monocytogenes*	[23]
S_aur1	SP:Q99TD3	*S. aureus* strain N315	[24]
S_aur2	SP:Q99UX5	*S. aureus* strain N315	[24]
S_aur3	SP:Q99UX3	*S. aureus* strain N315	[24]
S_aur4	SP:Q9KW67	*S. aureus* strain N315	[24]
B_anth	SP:Q9X358	*B. anthracis*, virulence plasmid PX01	[25]

*Protein accession numbers are shown as SP:xxxxxx for the SPTREMBL database, and as GB:xxxxxx for GenBank database. The corresponding sequences from *Staphylococcus aureus* strain Mu50 were identical to those from *S. aureus* strain N315 and were not included in the analysis.

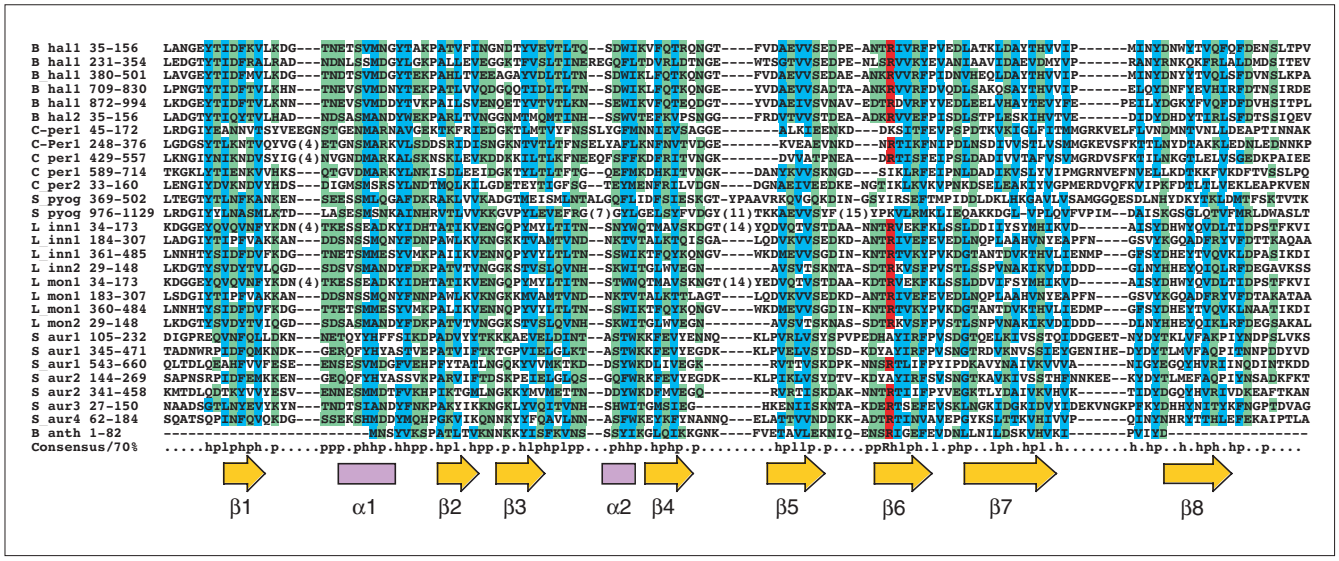

Figure 2
Multiple alignment of the occurrences of the NEAT domain, generated with the ClustalW program [28]. SMART [9,10], which identifies repeats using prospero [29,30], was used to search for domains in some sequences. The internal repeats detected in this manual analysis were used to generate subsequences that were used for building the first alignment. Then, we followed an iterative procedure by building a Hidden Markov Model (HMM) of the alignment and adding to the alignment significant hits from an HMM search [31,32] comparison of the HMM to the NCBI's nonredundant protein database. For the final HMM (derived from the alignment presented in this figure) no more similar sequences were detected below a standard *E*-value threshold (*E* = 0.001). The consensus in 70% of the sequences is reported below the alignment. Residue ranges are listed next to the protein code name.The letters h, l, and p indicate hydrophobic, aliphatic, and polar residues, respectively. Hydrophobic residues are highlighted in dark blue, polar residues in green, and a fairly conserved arginine (R in the consensus sequence) in red. Codes are the same as in Figure 1 and Table 1. The predicted secondary structure [8], mostly beta sheet, is displayed at the bottom of the figure. Although the B_anth sequence corresponds to a fragment of the domain, examination of the corresponding DNA sequence indicates that the actual translation product might extend further in both the amino- and carboxy-terminal directions.

carboxy-terminal transmembrane region (from SMART [9,10], using the Bioperl sigcleave module [11], based on [12]; and from THMM2 [13], respectively). In two *S. aureus* proteins (see Figure 3 for details) the transmembrane region prediction is over-ruled by the prediction of a carboxy-terminal motif that is typical of surface proteins of Gram-positive cocci. This motif consists of a fairly conserved hexapeptide followed by a hydrophobic anchor and two or three basic residues ([14]; detected using Pfam [15,16]). These features indicate a highly probable association between the proteins containing the NEAT domain and the membrane, and the signal peptide suggests that they are exported to the extracellular side of the membrane.

Phylogenetic analysis of the new domain

The phylogenetic tree constructed from the alignment of different domain occurrences (Figure 4) indicates a variety of independent duplication events. Some species contain up to four sequences with the domain, some only one. All repeats of B_hal1 (except one) cluster together; this indicates that one of the repeats of an ancestral *B. halodurans* sequence duplicated quickly (after divergence of *B. halodurans* from the other species displayed in the tree) into another three copies. The *C. perfringens* domains and the *S. aureus* domains are also the result of separated gene duplication and domain duplication events, as indicated by the clustering of the domains from these species. The clustering of the two *Listeria* species indicates no further duplication event in these species after their (recent) divergence.

Conclusions

Some protein domains have a highly variable copy number per protein, but they can exist as a single copy. This is in contrast to structural repeats (such as armadillo or leucine-rich repeats) that fold together and, by definition, never appear as a single copy [17]. Whereas structural repeats are related to DNA or protein binding, occasionally repeated domains can bind either large or small substrates; for example, Ca^{2+} (bound by C2, cadherin repeats, epidermal growth factor repeats), nucleotides (bound by zinc finger domains, LIM domains, and homeobox domains), or proteins (kazal inhibits serine proteases, ubiquitin domains in polyubiquitin bind target proteins to be degraded, PDZ domains bind polypeptides, nebulin repeats bind actin, immunoglobulins bind antigens, fibronectin 1 repeats bind fibrin and so on). (See the SMART server for further examples and references [9,10].) Accordingly, occasionally repeated domains are often involved in signaling or transcription regulation. A large copy number is used as a way of

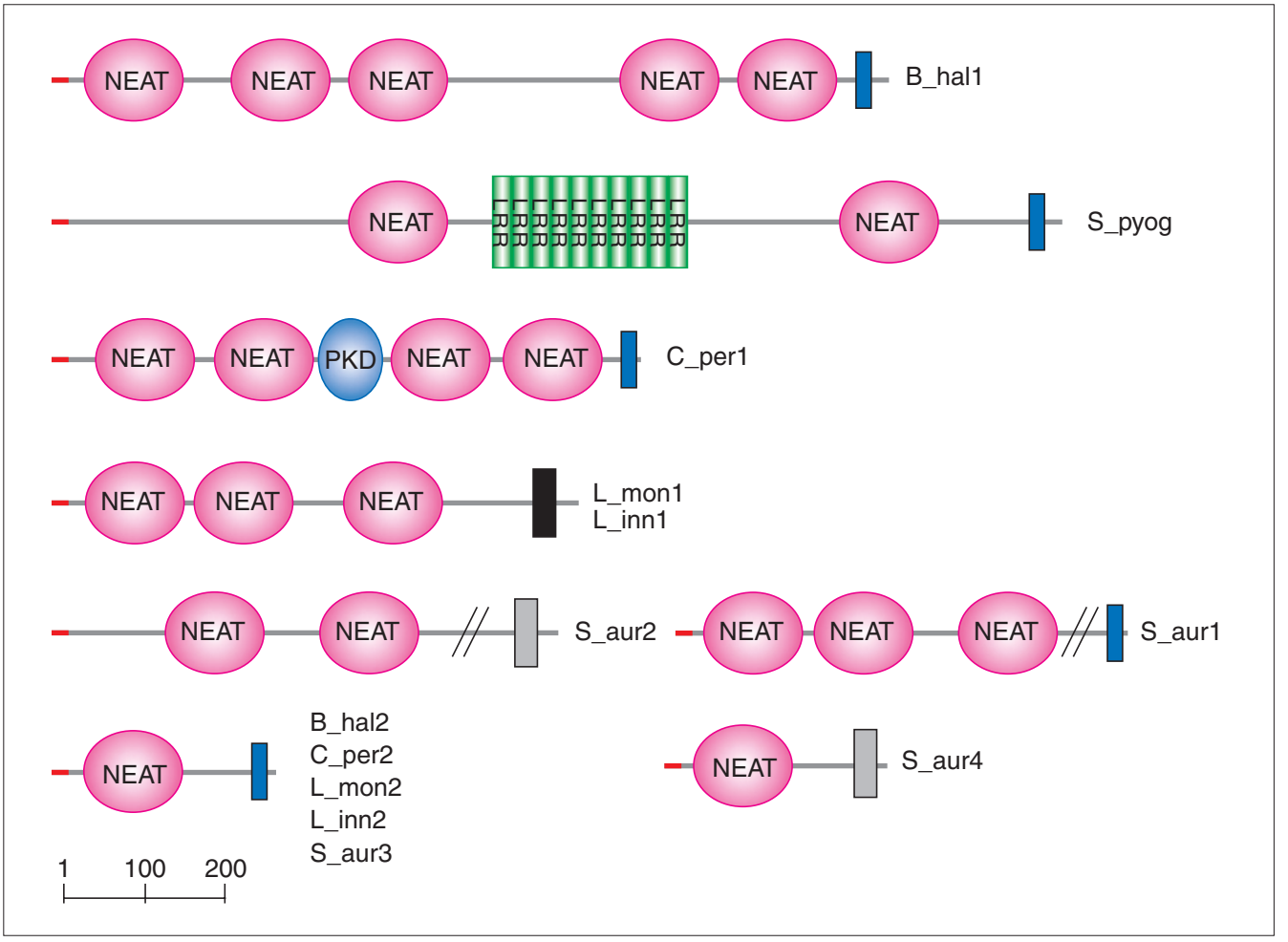

Figure 3
Modular arrangements of sequences containing the NEAT domain. Protein codes are the same as those shown in Figure 1 and Table 1. The red line indicates the signal peptide; the blue box represents a transmembrane helix; the gray box indicates the Gram-positive anchor as detected by Pfam [15,16]; the black box represents a hydrophobic carboxy-terminal anchoring domain proposed for two *Listeria* sequences [23]. PKD is the polycystic kidney disease domain (present in PKD1, chitinases, and collagenases, among others), and LRR stands for leucine-rich repeat (ten copies detected in S_pyog, using the program REP [33,34]). The scale bar indicates the length in amino acids.

increasing the effectiveness of the binding activity. This could be the case with the NEAT domain, which can be found as one single copy per sequence. In this respect, the NEAT domain appears to perform a binding function rather than a structural or an enzymatic one. Accordingly, the multiple alignment of the instances of the domain (Figure 2) indicates the lack of obvious conserved catalytic residues.

The NEAT domain appears to be associated with iron transport in several Gram-positive species (some of them pathogenic). Given its predicted extracellular location and its close association with the components of an iron transport system, one possible function of the NEAT domain is to be a receptor of the siderophore-iron complex. It would initiate a cascade upon detection of the substrate, ending in the expression of the components of the transporter in a system similar to that used in the induction of FecA [18]. Further evidence in this direction is given by recent experimental results for two of the NEAT-domain proteins from *S. aureus*, FrpA and FrpB (denoted here as S_aur4 and S_aur2, respectively), which were identified as cell wall proteins expressed under iron-restricted conditions [19].

The multiple duplication of this domain could reflect competition with an inhibitor. It could also be used for increasing bacterial sensitivity to the presence of the iron complex at very low substrate concentrations, in order to trigger the production of the corresponding transporter. The extracellular location of the domain, its association with a key process for bacterial survival, and its specificity to the group of pathogenic bacteria described, all make it a good candidate for developing a strategy against these pathogens.

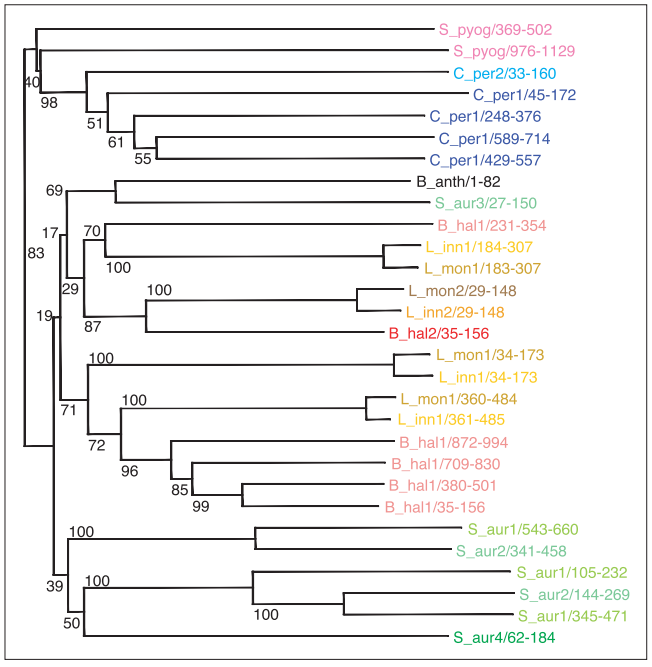

Figure 4
Phylogenetic tree of the domain instances generated from the multiple alignment shown in Figure 2. Bootstrapping values range from 0 to 100. The labels indicate the sequence and position of the repeat in the sequence. Domains from the same sequence have identical color (for example, all B_hal1 repeats are red). Domains from sequences of the same species have similar colors (for example, the *S. aureus* domains are colored in different hues of green).

References

1. Ratledge C, Dover LG: **Iron metabolism in pathogenic bacteria.** *Annu Rev Microbiol* 2000, **54:**881-941.
2. Staudenmaier H, van Hove B, Yaraghi Z, Braun V: **Nucleotide sequences of the *fecBCDE* genes and locations of the proteins suggest a periplasmic-binding-protein-dependent transport mechanism for iron(III) dicitrate in *Escherichia coli*.** *J Bacteriol* 1989, **171:**2626-2633.
3. Ozenberger BA, Nahlik MS, McIntosh MA: **Genetic organization of multiple *fep* genes encoding ferric enterobactin transport functions in *Escherichia coli*.** *J Bacteriol* 1987, **169:**3638-3646.
4. Burkhardt R, Braun V: **Nucleotide sequence of the *fhuC* and *fhuD* genes involved in iron (III) hydroxamate transport: domains in FhuC homologous to ATP-binding proteins.** *Mol Gen Genet* 1987, **209:**49-55.
5. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.
6. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
7. Huynen M, Snel B, Lathe W, Bork P: **Exploitation of gene context.** *Curr Opin Struct Biol* 2000, **10:**366-370.
8. Rost B, Sander C: **Combining evolutionary information and neural networks to predict protein secondary structure.** *Proteins* 1994, **19:**55-72.
9. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource.** *Nucleic Acids Res* 2002, **30:**242-244.
10. **SMART - Simple Modular Architecture Research Tool** [http://smart.embl-heidelberg.de/]
11. **Perldoc documentation for Bioperl Modules** [http://doc.bioperl.org]
12. von Heijne G: **A new method for predicting signal sequences cleavage sites.** *Nucleic Acids Res* 1986, **14:**4683-4690.
13. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL: **Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes.** *J Mol Biol* 2001, **305:**567-580.
14. Fischetti VA, Pancholi V, Schneewind O: **Conservation of a hexapeptide sequence in the anchor region of surface proteins from gram-positive cocci.** *Mol Microbiol* 1990, **4:**1603-1605.
15. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
16. **Protein families database of alignments and HMMs** [http://www.sanger.ac.uk/Pfam/]
17. Andrade MA, Perez-Iratxeta C, Ponting CP: **Protein repeats: structures, functions and evolution.** *J Struct Biol* 2001, **134:**117-131.
18. Enz S, Mahren S, Stroeher UH, Braun V: **Surface signaling in ferric citrate transport gene induction: interaction of the FecA, FecR, and FecI regulatory proteins.** *J Bacteriol* 2000, **182:**637-646.
19. Morrissey JA, Cockayne A, Hammacott J, Bishop K, Denman-Johnson A, Hill PJ, Williams P: **Conservation, surface exposure, and *in vivo* expression of the Frp family of iron-regulated cell wall proteins in *Staphylococcus aureus*.** *Infect Immun* 2002, **70**:2399-2407.
20. Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, Fuji F, Hirama C, Nakamura Y, Ogasawara N, Kuhara S, Horikoshi K: **Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*.** *Nucleic Acids Res* 2000, **28:**4317-4331.
21. Shimizu T, Ohtani K, Hirakawa H, Ohshima K, Yamashita A, Shiba T, Ogasawara N, Hattori M, Kuhara S, Hayashi H: **Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater.** *Proc Natl Acad Sci USA* 2002, **99:**996-1001.
22. Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, Primeaux C, Sezate S, Suvorov AN, Kenton S, *et al.*: **Complete genome sequence of an M1 strain of *Streptococcus pyogenes*.** *Proc Natl Acad Sci USA* 2001, **98:**4658-4663.
23. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, Berche P, Bloecker H, Brandt P, Chakraborty T, *et al.*: **Comparative genomics of *Listeria* species.** *Science* 2001, **294:**849-852.
24. Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K, Nagai Y, *et al.*: **Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*.** *Lancet* 2001, **357:**1225-1240.
25. Okinaka RT, Cloud K, Hampton O, Hoffmaster AR, Hill KK, Keim P, Koehler TM, Lamke G, Kumano S, Mahillon J, *et al.*: **Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes.** *J Bacteriol* 1999, **181:**6509-6515.
26. Snel B, Lehmann G, Bork P, Huynen MA: **STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene.** *Nucleic Acids Res* 2000, **28:**3442-3444.
27. **STRING - Search Tool for Recurring Instances of Neighbouring Genes** [http://www.bork.embl-heidelberg.de/STRING/]
28. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
29. Mott R: **Accurate formula for P-values of gapped local sequence and profile alignments.** *J Mol Biol* 2000, **300:**649-659.
30. **PROSPERO** [http://www.well.ox.ac.uk/rmott/ARIADNE/prospero.shtml]
31. Eddy SR: **Profile hidden Markov models.** Bioinformatics *1998,* **14**:755-763.
32. **HMMER 2.2** [http://hmmer.wustl.edu/]
33. Andrade MA, Ponting CP, Gibson TJ, Bork P: **Homology-based method for identification of protein repeats using statistical significance estimates.** *J Mol Biol* 2000, **298:**521-537.
34. **REP** [http://www.embl-heidelberg.de/~andrade/papers/rep/search.html]