**Open Journal of Genetics, 2012, 2, 18-22** OJGen
Published Online December 2012 (http://www.SciRP.org/journal/ojgen/)

# A bio-informatics study of the c25 cysteine protease family

**K.J. Cross, N.L. Huq, E.C. Reynolds**

Bio21 and Oral Health CRC, University of Melbourne, Melbourne 3010, Australia
Email: e.reynolds@unimelb.edu.au

Received 2012

## ABSTRACT

**The oral pathogen *Porphyromonas gingivalis* is recognized as one of the major aetiological agents of chronic periodontitis. The gingipains, which are the principal virulence factors of *P. gingivalis*, are multi-domain proteins containing an N-terminal C25 cysteine protease domain. We have conducted a bio-informatics study of the C25 cysteine protease domains and have identified related domains in over two thousand proteins from 739 organisms in 35 distinct phyla. Proteins having significant similarity to the gingipain C25 cysteine protease domain are also found in Gram +ve bacteria, Archaea, algae, higher fungi, and a wide variety of Eukaryotic species.**

**Keywords:** C25 Cysteine Proteases; Evolution; Substrate Preference

## 1. INTRODUCTION

Gingivitis is an inflammatory disease of the gum tissue. If not checked, the disease can progress to periodontitis leading to inflammation of the soft tissues surrounding the teeth, resorption of bone, and eventual loss of teeth. *Porphyromonas gingivalis*, is a major pathogen associated with chronic periodontitis in adults.

Gingipains were identified as the outer membrane, multi-domain virulence factors of the oral pathogen *Porphyromonas gingivalis* [1]. The N-terminal domain of the gingipains is a C25 cysteine protease domain. The evolution of the C25 cysteine protease family has been difficult to elucidate due to both the limited number of family members identified and their narrow distribution by species. The aim of this study was to undertake a detailed search for C25 protease-like sequences in public genome databases.

The MEROPS database [2] defines clan CD as containing families of proteases with either a protein fold or a sequence motif similar to those found in the caspase family (C14) and a histidine-cysteine catalytic dyad with the histidine located N-terminal to the cysteine. The catalytic histidine is usually in a histidine-glycine motif and is preceded by a block of hydrophobic residues. The catalytic cysteine is found predominantly in an alanine-cysteine motif and is preceded by a second block of hydrophobic residues.

Enzyme specificity in clan CD is determined primarily by the P1 residue of the substrate, which is normally an asparagine in family C13 (legumains), an aspartate in family C14 (caspases), and either an arginine or lysine in C11 (clostripains), C25 (gingipains), and C50 (separases), and a leucine in C80 (RTX toxin) [2]. The C25 (gingipain) specificity preference is based on the limited experimental data available for three proteins (RgpA, RgpB, and Kgp) from *Porphyromonas gingivalis*: RgpA and RgpB share greater than 97% sequence identity through their respective catalytic domains and display specificity toward arginine residues, while the more divergent Kgp displays specificity toward lysine residues.

Tertiary structures are only available for members of families C14 [3], C25 [4] and C80 [5]. These show α/β-proteins with a fold consisting of an α/β/α sandwich. The β-sheet contains six strands (in the order 213456) with strand 6 anti-parallel to the rest. The fold is believed to be unique to members of clan CD [2]. Other protein families are included in clan CD because of the conservation of motifs around the catalytic residues [6].

## 2. MATERIALS AND METHODOLOGY

### 2.1. Fugue Runs

We used a "pattern initiated hit and search" strategy to identify possible C25 domain sequences in the 'nr' database [7]. In brief, a preliminary alignment of C25 cysteine protease domain sequences was used to develop a regular expression that described the key features of the alignment. The regular expression was then used to pare down the number of sequences to be considered in the second step of the process.

Sequences that passed the regular expression were aligned against the RgpB sequence using Fugue [8] taking into account the structural preferences of residues. Sequences with a Z-score greater than 5.94 were accepted for further study. Fugue alignments were

performed using the structurally annotated sequence (hs1cvra) for RgpB available from the HOMSTRAD database [9] and the Fugue program [8]. The sequence and structure of RgpB used throughout this paper is that of Eichinger *et al*. [4].

### 2.2. Sequence Alignments and Analysis

T-Coffee [10] was used in the 'espresso' mode to generate an initial (master) alignment of the C25 domain sequences identified using Fugue. MAFFT [11] was used to generate sequence alignments of the cysteine protease domains identified using PSI-BLAST [12]. To ensure consistency, all alignments were performed using the fft option and 're-treeing' twice. The quality of the hits identified as possible C25 cysteine protease domains was assessed using sequence alignments and the Shannon information content of those alignments [13]. A Tcl/Tk script was developed that populated the EPS template used by WebLogo and allowed selected columns from the sequnce alignments to be plotted.

### 2.3. PSI-BLAST Runs

Preliminary runs of PSI-BLAST were performed using a single C25 domain sequence (RgpB) and varying the 'inclusion threshold' and maximum 'E value'. Two runs were performed with a maximum E-value of 1e-3 and inclusion threshold values of 5e-4 and 5e-3. A further two runs were performed with an inclusion threshold of 5e-4 and maximum E-values of 1e-2 and 1e-3.

Two production runs were performed. The first using a master sequence alignment of 103 protein sequences (MSA) identified by Fugue as being C25 domains. The inclusion threshold was set at 5e-4, the maximum E-value of 1e-4, and 6 iterations of the PSI-BLAST algorithm were performed. The master sequence index was incremented in consecutive runs so that a PSI-BLAST was performed for each of the C25 sequences identified by Fugue. The PSI-BLAST data was aggregated into an sqlite database for further analysis. The second run used a curated subset of the sequences identified in the third iteration of the first run with 336 sequences (MSA) and three PSI-BLAST iterations were performed and analysed as for the first run.

## 3. RESULTS AND DISCUSSION

### Identification of Seed Sequences

A total of 103 sequences were identified as possible C25 cysteine protease domains using the Fugue-filter criterion. They constitute a subset of the 185 sequences identified in the Pfam database [14]. However, at least six of the sequences identified by Pfam as C25 cysteine protease sequences appear to lack the catalytic cysteines. These non-cysteine protease sequences are A9B8P8_HERA2,A4BYP5_9FLAO,E1K598_9EURY, P96966_PORGI, E1K599_9EURY, and A7BN31_9GAMM.

PSI-BLAST uses two parameters to control the search for related sequences. The 'expectation value' (E-value) threshold determines whether a hit is accepted or not, smaller values are associated with increased significance. The 'inclusion threshold' is the maximum expectation value for a hit to be used to calculate the PSSM (Position Sensitive Substitution Matrix), the matrix describing the probability of particular mutations occurring at specific locations in the protein sequence.

In preliminary experiments, we demonstrated that changing the E-value threshold from 1.0e-3 to 1.0e-2, while keeping the inclusion threshold fixed at 5.0e-4, increased the number of hits from 327 to 332 after five iterations starting from the RgpB C25-domain sequence. Changing the 'inclusion threshold' had no effect on the number of hits located. Having demonstrated that the PSI-BLAST algorithm was fairly insensitive to these parameters at these levels (results not shown), we used a significance level of 5.0e-4 for the inclusion threshold and 1.0e-4 for the E-value threshold to minimize the risk of incorporating misassigned sequences.**Figure 1** shows

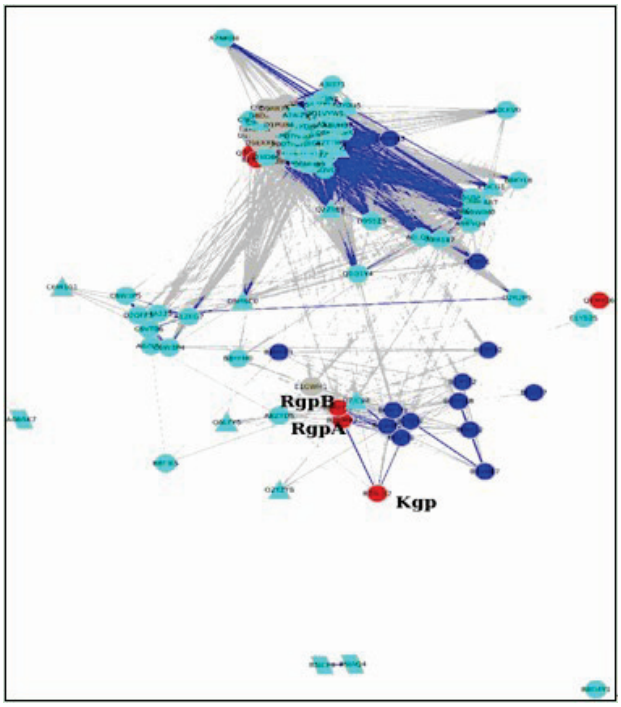


**Figure 1:** Schematic representation of the 'BLAST connectivity' of sequences in the 'seed' set of C25 sequences. Heavier lines represent smaller, more significant expectation values in the range $1e^{-10}$ to $1e^{-420}$. Note the location of the archetypical RgpA, RgpB, and Kgp C25 sequences on the periphery of the cluster, and the tight cluster of sequences primarily associated with bacteria from the *Prevotella* phylum.

the 'BLAST connectivity' network for the C25 domains

 

identified by the Fugue-filter process: the plot was prepared using Cytoscape [15]. Heavier lines indicate smaller expectation values for the blast connectivity between the C25 domains (i.e. higher significance). All the connectivities shown in **Figure 1** have expectation values of 1e-10 or less and can be considered as highly significant. **Figure 1** emphasizes the fact that the C25 domains of RgpA, RgpB, and Kgp (the archetypical C25-proteins) are in fact on the edge of a cluster of proteins that should be thought of as having typical C25 domain sequences.

As shown in **Figure 2** there is a rapid increase in the number of sequences identified after the third iteration of PSI-BLAST accompanied by a rapid decline in the information content of the alignments whether considered as a sum over all positions in the alignment, or as information per position as shown. This behaviour could be consistent with 'contamination' of the PSSM used by PSI-BLAST. The observation that the rapid increase in the number of hits occurred even after the removal of suspected non-gingipain sequences in the second, production PSI-BLAST run suggests that many of these additional hits are to significantly related sequences.

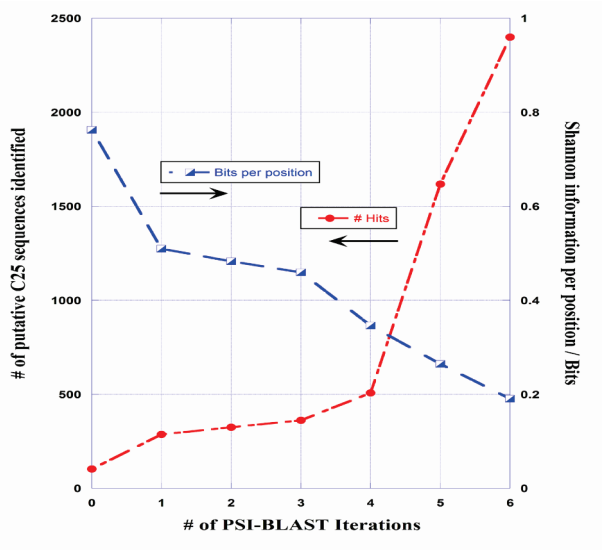


**Figure 2.** Number of unique sequences identified and the information content per position in the sequence alignments as a function of PSI-BLAST iteration using the 103 C25-sequences identified by Fugue as the 'seed' set.

The putative C25 cysteine protease domains were curated by inspection of the alignments. For example, inspection of the alignment of 362 sequences identified at iteration 3 of the first PSI-BLAST run identified 44 sequences that either lacked a conserved cysteine at the catalytic site, or had residues such as valine or proline aligned with the catalytic histidine, or had a cluster of three bulky residues aligned with the 'GHG'-motif. These sequences were removed from the alignment and their GI numbers added to a 'black list' of sequences to be excluded from future PSI-BLAST runs. The information per position increased from 0.45954 ± 0.00029 to 0.49367 ± 0.00034 after removal of the 'bad' sequences (compared to 0.50987 ± 0.00038 at the first iteration).

**Figure 3** shows a 'WebLogo'-style representation [16] of those columns that have fewer than 40% gaps in the alignment of the 2,333 proteins identified as being related to C25 domains in this work. The gaps between conserved portions (note the indices are discontinuous) are consistent with a family of proteins having a conserved core connected by more variable regions.

Table 1 summarizes the progress of the search for C25 cysteine protease domains. The number of Archaeal and Bacterial proteins identified increases at each stage of the search process, as does the number of Eukaryotic species. This suggests the families of proteins within Clan CD are significantly overlapped in sequence space.

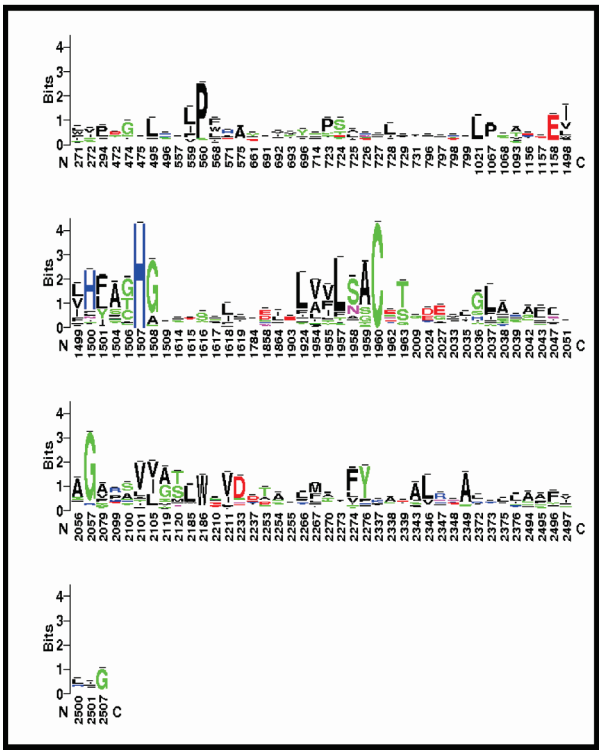


**Figure 3.** A 'WebLogo'-style representation of the mafft alignment of the 2,333 proteins identified in the final, curated psi-blast data set. To compress the graphic, only those columns with at most 40% gaps are represented as a consequence the column indices labelling the x-axis are discontinuous.

**Figure 4** emphasizes the structural similarities between RgpB (C25), Yca1 a meta-caspase from yeast (C14B), the RTX toxin from *V. cholera*e (C80), and human caspase-7 (C14). Not only is there a strong topological

 

**Table 1.** Summary of the number of different species and proteins organized by phyla at various stages of the search for C25 cysteine protease domains. The 'Seed Set' consists of the 103 proteins confirmed by Fugue alignment to be 'C25 cysteine protease domains'. The '1st Round' proteins are those identified in three iterations of PSI-BLAST starting from the 'Seed Set', but excluding sequences as described in the text. The 'Final' set represents the proteins identified after a further three iterations from the "1st Round" set. The figure in brackets represents the number of proteins.

| Groups | Phylum | Seed Set | 1st Round | Final |
|---|---|---|---|---|
| **Archaea** | | | | |
| | Crenarchaeota | | 2 (4) | 1 (1) |
| | Euryarchaeota | 2 (2) | 7 (19) | 17 (26) |
| | Korarchaeota | | 1 (1) | 1 (1) |
| | Thaumarchaeota | | | 1 (2) |
| **Bacteria** | | | | |
| | Acidobacteria | | 1 (1) | 6 (18) |
| | Actinobacteria | | 4 (4) | 89 (208) |
| | Aquificae | | | 1 (1) |
| | Bacteroidetes | 56 (65) | 131 (210) | 149 (333) |
| | Caldiserica | | 1 (1) | 1 (1) |
| | Chlamydiae | | | 1 (13) |
| | Chlorobi | 1 (1) | 1 (2) | 6 (11) |
| | Chloroflexi | 6 (7) | 8 (19) | 10 (73) |
| | Cyanobacteria | | 3 (3) | 53 (677) |
| | Deinococcus-Thermus | | | 1 (1) |
| | Firmicutes | | 1 (1) | 12 (13) |
| | Fusobacteria | | | 1 (1) |
| | Gemmatimonadetes | | | 1 (1) |
| | Ignavibacteria | | 2 (3) | 2 (4) |
| | Nitrospirae | | | 2 (7) |
| | Planctomycetes | 3 (3) | 5 (5) | 6 (14) |
| | Proteobacteria | 5 (8) | 31 (42) | 176 (317) |
| | Spirochaetes | | 13 (13) | 18 (19) |
| | Verrucomicrobia | 1 (1) | 1 (2) | 2 (2) |
| **Fungi** | | | | |
| | Ascomycota | | 2 (4) | 76 (184) |
| | Basidiomycota | | | 21 (248) |
| **Eukaryota** | | | | |
| | Arthropoda | | | 3 (3) |
| | Bacillariophyta | | | 2 (3) |
| | Chordata | | | 29 (33) |
| | Cnidaria | | | 1 (24) |
| | Echinodermata | | | 1 (1) |
| | Platyhelminthes | | | 1 (1) |
| | Porifera | | | 1 (2) |
| **Viridiplantae** | | | | |
| | Chlorophyta | | | 2 (2) |
| | Streptophyta | | | 13 (30) |
| **Brown algae** | | | | |
| | Phaeophyceae | | | 1 (3) |

similarity as shown by the order and orientation of the β-strands forming the core of the proteins, but the catalytic histidine and cysteine are consistently two residues

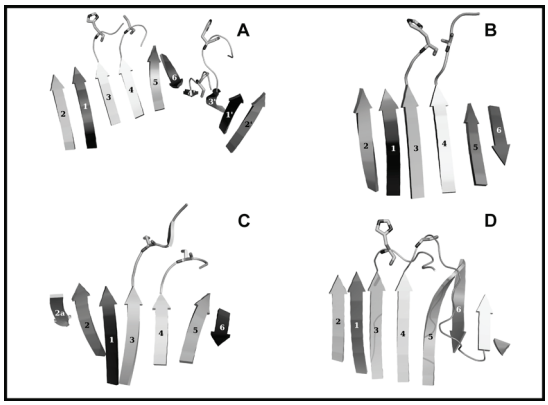


**Figure 4.** Cartoon representations of the β-strand core of RgpB from *Porphyromonas gingivalis* (PDB:1cvr) a C25 cysteine protease (A), caspase-7 from *Homo sapiens* (PDB:1k88) a C14 cysteine protease (B), the RTX-toxin from *Vibrio cholerae* (PDB:3gcd) a C80 cysteine protease (C), and Yca1 a metacaspase from *Saccharomyces cerevisiae* (PDB:4f6o) a C14B cysteine protease (D). The location of the catalytic histidine and cysteine shown in stick form, are always located two residues C-terminal of the final residues in the third and fourth β-strands respectively.

C-terminal to the last residues of the 3rd and 4th β-strands respectively. Given this strong structural similarity, it is not surprising that C25 domain proteins display a marked sequential similarity with other Clan CD proteins.

The sequence specificity of the C25 proteases is determined by the residues that line the 'S1' specificity pocket. As shown in **Figure 5** there is very little sequence conservation in this region apart from the 'GHG'-motif and the catalytic cysteine. For example, in RgpB the side-chain of Asp163 hydrogen bonds to the guanidine group of the arginine substrate whereas in Kgp the analogue of Asp163 is a threonine and it is probably Asp516 that is the predominant hydrogen-bonding partner to the lysine substrate. Inspection of the alignment of the putative gingipain sequences identifies three sequences that had previously been annotated as either "propeptide peptidase C25" (GI: 373458037) or "peptidase C25" (GI:326279641 and 307565663) from *Caldithrix abyssi*, *Odoribacter splanchnicus*, and *Prevotella amnii* [7] respectively, where Thr209 of RgpB is replaced by an arginine residue. Thr209 is located near the bottom of the 'S1' specificity pocket, and an arginine residue would have its side-chain extend to near the top of the specificity pocket suggesting that these three bacterial gingipains have a caspase-like specificity toward aspartate (or possibly glutamate) residues N-terminal to the cleavage point. In, for example, RgpB a short-chain residue near the bottom of the specificity pocket is used to 'recognize' the long-chain of the arginine substrate. The three proteins identified here, appear to use a long-chain arginine at a neighbouring location to 'recognize' a short-chain substrate. These specific examples underscore the wide range of substrate specificities sug-

 

gested by **Figure 5**.

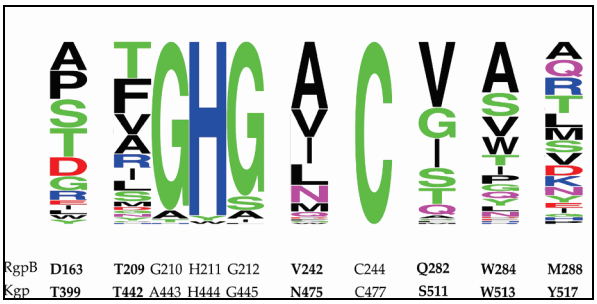


**Figure 5**. A plot of the frequency of occurrence of various residues that align with the 'S1'-pocket residues of RgpB in the 'seed' set of 103 C25 cysteine protease sequences. As shown in Table 1, these are Bacterial or Archaeal proteins. These residues determine the substrate specificity of the various C25 proteases. The low overall conservation suggests a broad range of C25 substrate preferences.

## 4. CONCLUSION

The bacterial C25 cysteine proteases share significant sequential and structural similarity with other Clan CD cysteine proteases. The number of identified bacterial and archaeal sequences increases at each stage of the search procedure as seen in Table 1, while the number of sequences associated with other phyla increases dramatically in the final round of the search.

The lack of sequence conservation in the 'S1'-binding site argues for a wide-range of substrate specificities among the C25 cysteine proteases further blurring the distinctions between the various protease families within the Clan CD proteases.

## 5. ACKNOWLEDGEMENTS

We acknowledge funding from the Oral Health CRC and NH&MRC.

## REFERENCES

[1] Carlsson, J., B.F. Herrmann, J.F. Hofling, and G.K. Sundqvist. (1984) Degradation of the human proteinase inhibitors alpha-1-antitrypsin and alpha-2-macroglobulin by Bacteroides gingivalis. *Infection and Immunity*. **43,** 644-648.

[2] Rawlings, N.D., A.J. Barrett, and A. Bateman. (2010) MEROPS: the peptidase database. *Nucleic Acids Research*. **38,** D227-D233. doi:10.1093/nar/gkp971

[3] Walker, N.P.C., R.V. Talanian, K.D. Brady, L.C. Dang, N.J. Bump, *et al.* (1994) Crystal structure of the cysteine protease interleukin-1 beta-converting enzyme: a (p20/p10)2 homodimer. *Cell*. **78,** 343-352. doi:10.1016/0092-8674(94)90303-4

[4] Eichinger, A., H.G. Beisel, U. Jacob, R. Huber, F.J. Medrano, *et al.* (1999) Crystal structure of gingipain R: an Arg-specific bacterial cysteine proteinase with a caspase-like fold. *EMBO Journal*. **18,** 5453-5462. doi:10.1093/emboj/18.20.5453

[5] Lupardus, P.J., A. Shen, M. Bogyo, and K.C. Garcia. (2008) Small molecule-induced allosteric activation of the Vibrio cholerae RTX cysteine protease domain. *Science*. **322,** 265-8. doi:10.1126/science.1162403

[6] Chen, J.M., N.D. Rawlings, R.A. Stevens, and A.J. Barrett. (1998) Identification of the active site of legumain links it to caspases, clostripain and gingipains in a new clan of cysteine endopeptidases. *FEBS Letters*. **441,** 361-365. doi:10.1016/S0014-5793(98)01574-9

[7] nr database. [cited 2012 September 3]; Available from: ftp://ftp.ncbi.nih.gov/blast/db.

[8] Shi, J., T.L. Blundell, and K. Mizuguchi. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology*. **310,** 243-257. doi:10.1006/jmbi.2001.4762

[9] Mizuguchi, K., C.M. Deane, T.L. Blundell, and J.P. Overington. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Science*. **7,** 2469-2471. doi:10.1002/pro.5560071126

[10] Notredame, C., D.G. Higgins, and J. Heringa. (2000) T-Coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology*. **302,** 205-217.

[11] Katoh, K., K. Misawa, K. Kuma, and T. Miyata. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acid Research*. **30,** 3059-3066

[12] Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*. **10,** 421. doi:10.1186/1471-2105-10-421

[13] Shannon, C.E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*. **27,** 379–423.

[14] [14] Côté, R.G., P. Jones, L. Martens, S. Kerrien, F. Reisinger, *et al.* (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*. **8,** 401. doi:10.1186/1471-2105-8-401

[15] [15] Smoot, M.E., K. Ono, J. Ruscheinski, P.L. Wang, and T. Ideker. (2010) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. **27,** 431-2. doi:10.1093/bioinformatics/btq675

[16] Crooks, G.E., G. Hon, J.M. Chandonia, and S.E. Brenner. (2004) WebLogo: a sequence logo generator. *Genome Research*. **14,** 1188-1190. doi:10.1101/gr.849004

 