OASYS: Orthology Assignment based on Synteny and Sequence Information

Tsuyoshi Hachiya* **Yasubumi Sakakibara***
hacchy@dna.bio.keio.ac.jp yasu@bio.keio.ac.jp

* Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kouhoku-ku, Yokohama, Kanagawa 223-8522, Japan

September 3, 2009

The OASYS algorithm

In this paper, we first propose a novel method to quantify the extent of gene order conservation, which can be applied to pairwise genome comparison. This quantification makes it simple to extract positonal orthologs from arbitrary many-to-many orthology assignments. Subsequently, the method for detecting positional orthologs without other resources of many-to-many orthology assignments is also proposed by integrating information about protein sequence conservation and gene order conservation on the basis of stochastic modeling and probabilistic theory. Both algorithms are implemented as a C++ program, named OASYS (Ortholog Assigner based on SYnteny and Sequence information). For the purpose of demonstrating the usefulness of OASYS, we compared the performance of OASYS in terms of the detection of positional orthologs with the existing algorithms, and subsequently, the relation between protein sequence evolution and gene order conservation is examined. Finally, we validated the possibility of using positional orthologs as alternatives of functional orthologs.

1 Weighted number of neighboring seed orthologs

1.1 Pre-computation

OASYS quantifies the extent of gene order conservation by a novel measure named weighted number of neighboring seed orthologs (WNNSO). The calculation of the WNNSO values requires homologous gene pairs to be pre-computed. Given all protein sequences encoded in two genomes A and B, the detection of the homologous gene pairs starts with the calculaiton of the pairwise sequence similarity scores by reciprocally using the BLASTP program [Altschul *et al.*, 1990]. Since the BLASTP program occasionally reports asymmetric scores, and the asymmetricity could cause problems in later steps [Remm *et al.*, 2001], all pairwise sequence scores are averaged.

Next, spurious BLAST matches are filtered out on the basis of two criteria. At first, matches whose bit score is less than SCORE_CUTOFF are filtered out. Second, orthologous genes are expected to maintain the homology over the majority of their length [Remm *et al.*, 2001]. Thus, matches, in which the length of the matching segments is less than OVERLAP_CUTOFF% of the length of the longer sequence, are filtered out. The remaining matches are regarded as homologous gene pairs, and used in later steps. The two parameters, SCORE_CUTOFF and OVERLAP_CUTOFF, are user adjustable parameters. The default value for the SCORE_CUTOFF parameter is set at 50 bits, and the default value for the OVERLAP_CUTOFF parameter is set at 50%.

1.2 Calculation of WNNSO values

[Bandyopadhyay *et al.*, 2006] identifies functional orthologs on the basis of the concept that a protein and its functional ortholog are likely to interact with proteins in their respective networks that are themselves functional orthologs. Analogously, OASYS identifies positional orthologs on the basis of the concept that a gene and its positional ortholog are likely to be located on their respective chromosomal positions that are diagonally proximate to themselves positional orthologs.

The calculation of the WNNSO values starts with the detection of putative orthologs. Putative orthologs are simply detected by the reciprocal best hit (RBH) method, that is, reciprocal best similarity pairs in terms of bit scores are detected as putative orthologs [Rivera *et al.*, 1998, Hirsh and Fraser, 2001, Jordan *et al.*, 2002]. We call the putative orthologs '*seed orthologs*', and the homologous gene pairs that are not identified as putative orthologs '*non-seed homologs*'.

Second, the diagonal proximities between homologous gene pairs and the seed orthologs are computed on the basis of the matrix representation of gene positions. Let $\mathbf{A}$ be a set of genes encoded by the genome A, $\mathbf{A}^k$ be a set of genes located on the k-th chromosome of the genome A, and a_i^k be the i-th gene located on the k-th chromosome. We assume without loss of generality that the elements in $\mathbf{A}^k$ are sorted in order of increasing start position along the k-th chromosome. Regarding genome B, $\mathbf{B}$, $\mathbf{B}^l$, and b_j^l are similarly defined. Then, a homologous gene pair (a_i^k, b_j^l) is represented as an element of a $|\mathbf{A}^k| \times |\mathbf{B}^l|$ matrix, in which a homologous gene pair (a_i^k, b_j^l) corresponds to a point (i, j). If two gene pairs $h_m = (a_i^k, b_j^l)$ and $h_{m'} = (a_{i'}^{k'}, b_{j'}^{l'})$ are *collinear*, a special distance function named diagonal pseudo distance (DPD) [Vandepoele *et al.*, 2002] is used to define the distance between the two gene pairs:

$$\mathrm{DPD}(h_m, h_{m'}) = 2\max(|i - i'|, |j - j'|) - \min(|i - i'|, |j - j'|). \tag{1}$$

If two gene pairs are not collinear, the distance is defined as infinity. The definition of the 'collinearity' can be found in the section 1.3.

Finally, the WNNSO value is computed for each homologous gene pair by counting the number of the seed orthologs near the homologous gene pair with weights that decrease with increasing the diagonal pseuso distance. Let $\mathbf{S}$ be a set of seed orthologs. Then, the WNNSO value for a homologous gene pair h_m is given by

$$\mathrm{WNNSO}(h_m|\mathbf{S}) = \sum_{h_{m'} \in \mathbf{S}} \mathrm{Weight}(h_m, h_{m'}) \tag{2}$$

$$\begin{aligned} &\mathrm{Weight}(h_m, h_{m'}) \\ &= \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{\mathrm{DPD}(h_m, h_{m'})}{2\sigma^2}) & \text{when } \mathrm{DPD}(h_m, h_{m'}) \leq \mathrm{Cut\_dpd} \\ 0 & \text{otherwise ,} \end{cases} \end{aligned} \tag{3}$$

where σ and Cut_dpd are user-defined parameters. σ controls the degree of the decrease of the weight value with increasing DPD, and Cut_dpd represents the threshold for DPD value. Note that the weight between non-collinear gene pairs becomes zero. Fig. 1 shows the result of applying the above DPD, weight and WNNSO function on a hypothetical example, and Fig. S1 shows the result of applying the weight function with various values of σ parameter, in which the effect of σ parameter can be observed.

1.3 Collinearity

If a gene x_i^k is located on the forward strand of the k-th chromosome, we denote $x_i^k.strand = 1$. If x_i^k is located on the reverse strand, we denote $x_i^k.strand = -1$. $y_j^l.strand$ is similarly defined. Then, the *sign* of a gene pair $h_m = (x_i, y_j)$ is defined as $h_m.sign = x_i.strand \times y_j.strand$. OASYS defines that

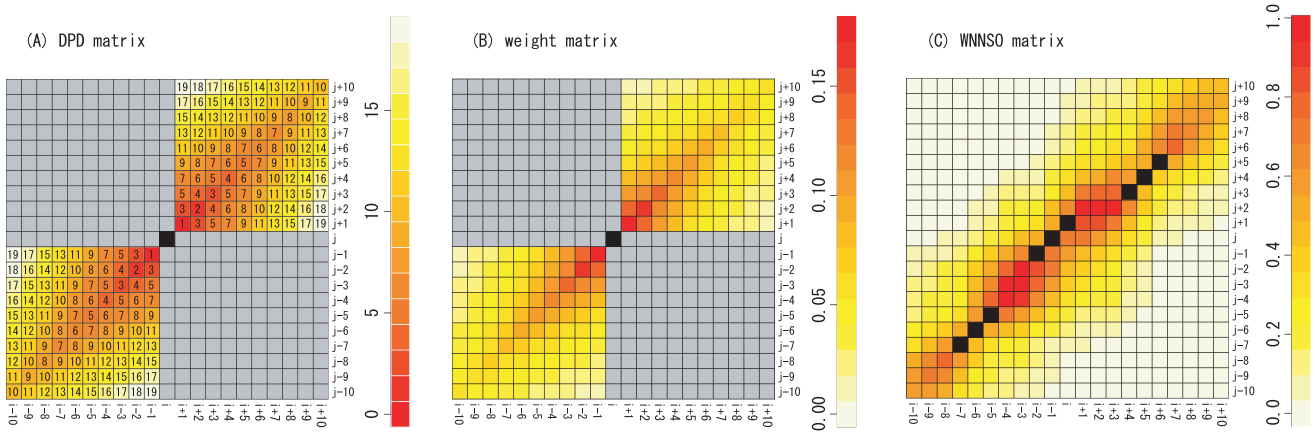

Figure 1: Graphical representation of the DPD, weight and WNNSO function. (A) DPD function. Given a homologous gene pair $h_m = (x_i, y_j)$ (represented as the central element colored by black), the color of a element $h_{m'} = (i', j')$ in the matrix represents the degree of the value of $DPD(h_m, h_{m'})$. Here, we assume that $h_m.sign = 1$ and $h_{m'}.sign = 1$. Positive integer shown in each element is the DPD value. Gray-colored elements in the matrix correspond to the gene pairs that are not collinear with the gene pair (x_i, y_j). The DPD value in these elements is defined as infinity. (B) Weight function. The color of a element $h_{m'} = (i', j')$ in the matrix represents the degree of the value of $\text{Weight}(h_m, h_{m'})$. When computing the weight values, the value of σ parameter was set at 2.0, and the value of cut_dpd parameter was set at 20. Regarding gray-colored elements, the value of $\text{Weight}(h_m, h_{m'})$ is computed as zero. (C) WNNSO function. Given a set of seed orthologs $\mathbf{S}$ (represented by elements colored by cyan), the color of each element $h_{m'}$ represents the degree of the value of $\text{WNNSO}(h_{m'}|\mathbf{S})$.

two gene pairs $h_m = (x_i^k, y_j^l)$ and $h_{m'} = (x_{i'}^{k'}, y_{j'}^{l'})$ are collinear if the following conditions are satisfied:

$$\begin{aligned} &k = k', \quad l = l', \quad i \neq i', \quad j \neq j', \\ &h_m.sign = h_{m'}.sign, \quad \frac{j - j'}{i - i'} \times h_m.sign > 0. \end{aligned} \tag{4}$$

2 Reduction from many-to-many orthology assignment to one-to-one orthology assignment

Suppose that we are given a set of many-to-many orthology assignments (i.e. a set of ortholog groups) calculated by arbitrary sequence-based algorithm. Then, after the pre-computation of homologous gene pairs, the extraction of one-to-one orthology assignments in terms of positional orthologs is performed based on the following steps:

1. Detect seed orthologs.

2. Perform the following sub-steps for each ortholog group.

   (a) Calculate WNNSO values for all pairs of genes in different species, which are included in the current ortholog group.

   (b) Detect positional orthologs by applying the RBH method to the set of genes included in the current ortholog group. In this RBH computation, OASYS uses the WNNSO values instead of sequence similarity scores. Of the RBH pairs in terms of WNNSO value, the RBH pairs whose WNNSO value is greater than zero are detected as positional orthologs.

Note that this algorithm would detect no positional orthologs from one ortholog group if all pairs of genes in the ortholog group have the WNNSO value of zero.

3 Orthology assignment without other resources

3.1 Detection of main orthologs

After the pre-computation of homologous gene pairs, the detection of positional orhtologs is performed based on the following steps:

1. Detect seed orthologs.
2. Calculate a WNNSO value for each homologous gene pair.
3. Model probability densities of the bit scores and the WNNSO values. The probability density functions used in OASYS are described in the section 3.2.
4. Calculate the integrated conservation score, which takes into account the gene order conservation as well as the protein sequence conservation, for each homologous gene pair. The scoring scheme used in OASYS is described in the section 3.3.
5. Detect main orthologs. The detection of the main orthologs is performed on the basis of the reciprocal best hit (RBH) approach. OASYS uses the integrated score in the RBH computation, while traditional RBH method simply uses the bit score or E-value.

3.2 Probability density functions

In order to distinguish between main orthologs and paralogs, OASYS takes advantage of the difference in the extent of the gene order conservation between main orthologs and paralogs. For this purpose, OASYS assumes that the probability density of the WNNSO values for main orthologs can be approximated by that for seed orthologs. It is also assumed that the probability density of the WNNSO values for paralogs can be approximated by that for non-seed homologs.

In addition to the difference in the extent of gene order conservation, OASYS also makes use of the difference in the extent of protein sequence conservation between main orthologs and paralogs. As in the case of the WNNSO values, OASYS assumes that the probability density of the bit scores for main orthologs (for paralogs) can be approximated by that for seed orthologs (for non-seed homologs).

In total, OASYS models four probability densities; (i) the probability density of the WNNSO values for main orthologs, (ii) the probaility density of the bit scores for main orthologs, (iii) the probability density of the WNNSO values for paralogs, and (iv) the probaility density of the bit scores for paralogs. Each of the four probability densities is modeled by either of two probability density functions (pdfs), namely the *one-sided generalized Gaussian* (OGG) pdf and the *asymmetric generalized Gaussian* (AGG) pdf. As shown later, the former can represent wide range of decreasing functions, and the later can represent wide range of unimodal functions. The choise of the model is performed on the basis of the Akaike information criteria [Akaike, 1974]. Supplementary Figs. S2 and S3 show that our model is well fitted to each data set. A detailed description about the model selection can be found in 'Model selection (S3.3)' in Supplementary Materials.

One-sided generalized Gaussian distribution. The generalized Gaussian (GG) distribution proposed in [Miller and Thomas, 1972] is given by

$$P_{\mathrm{gg}}(x;\mu,\sigma,p) = \begin{cases} \frac{p\gamma}{2\Gamma(\frac{1}{p})}\exp(-\gamma^p(\mu-x)^p) & \text{when } x<\mu \\ \frac{p\gamma}{2\Gamma(\frac{1}{p})}\exp(-\gamma^p(x-\mu)^p) & \text{when } x\geq\mu, \end{cases} \tag{5}$$

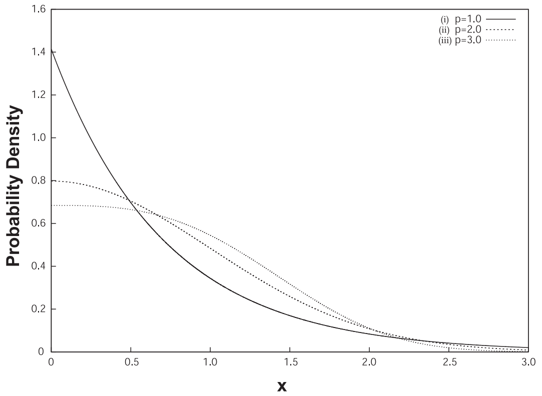

Figure 2: One-sided generalized Gaussian pdf. Three curves shown in this figure have the following parameter values; (i) $\mu = 0$, $\sigma^2 = 1$, and $p = 1.0$, (ii) $\mu = 0$, $\sigma^2 = 1$, and $p = 2.0$, (iii) $\mu = 0$, $\sigma^2 = 1$, and $p = 3.0$.

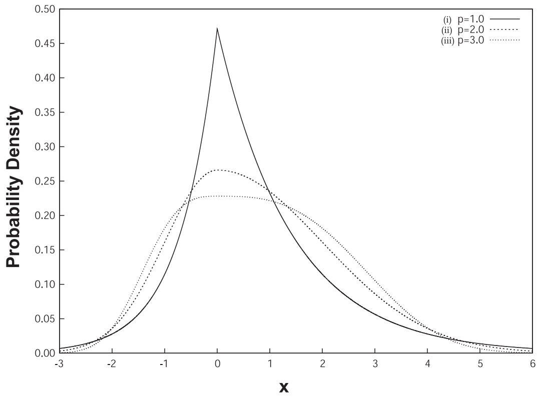

Figure 3: Asymmetric generalized Gaussian pdf. Three curves shown in this figure have the following parameter values; (i) $\mu = 0$, $\sigma_l^2 = 1$, $\sigma_r^2 = 2$ and $p = 1.0$, (ii) $\mu = 0$, $\sigma_l^2 = 1$, $\sigma_r^2 = 2$ and $p = 2.0$, (iii) $\mu = 0$, $\sigma_l^2 = 1$, $\sigma_r^2 = 2$ and $p = 3.0$.

where $\gamma = \frac{1}{\sigma}\sqrt{\frac{\Gamma(\frac{3}{p})}{\Gamma(\frac{1}{p})}}$ and $\Gamma(\bullet)$ is the gamma function. In this model, μ, σ^2, and p denote the mean, variance, and decay rate (also referred to as shape parameter) of the pdf, respectively. We modify Eq. (5) and define the one-sided generalized Gaussian (OGG) distribution, which is given by

$$P_{\text{ogg}}(x;\mu,\sigma,p) = \begin{cases} 0 & \text{when } x < \mu \\ \frac{p\gamma}{\Gamma(\frac{1}{p})}\exp(-\gamma^p(x-\mu)^p) & \text{when } x \geq \mu. \end{cases} \tag{6}$$

Note that μ in Eq. (6) is not the mean of the OGG pdf but a location parameter. For $x \geq \mu$, the pdf is a decreasing function of x. As shown in Fig. 2, the OGG family of distributions can represent wide range of decreasing functions by changing the shape parameter p.

Suppose that we are given a data set of observations of scalar values $\mathbf{x} = \{x_1, \ldots, x_N\}$ and that $x_i \geq \mu$ for $1 \leq i \leq N$. Then, the log likelihood function of the OGG pdf is given by

$$\ln L_{\text{ogg}} = N \ln\left(\frac{p\gamma}{\Gamma(\frac{1}{p})}\right) - \sum_{i=1}^{N} \gamma^p (x_i - \mu)^p. \tag{7}$$

We can optimize the parameters in the OGG model so as to maximize Eq. (7). For details, see 'Fitting to an OGG distribution (S3.1)' in Supplementary Materials.

Asymmetric generalized Gaussian distribution. The asymmetric generalized Gaussian (AGG) distribution proposed in [Tesei and Regazzoni, 1998] is given by

$$P_{\text{agg}}(x; \mu', \sigma_l, \sigma_r, q) = \begin{cases} \frac{q\gamma_a}{\Gamma(\frac{1}{q})} \exp(-\gamma_l^q(-x+\mu')^q) & \text{when } x < \mu' \\ \frac{q\gamma_a}{\Gamma(\frac{1}{q})} \exp(-\gamma_r^q(x-\mu')^q) & \text{when } x \geq \mu', \end{cases} \tag{8}$$

where $\gamma_a = \frac{1}{\sigma_l+\sigma_r}\sqrt{\frac{\Gamma(\frac{3}{q})}{\Gamma(\frac{1}{q})}}$, $\gamma_l = \frac{1}{\sigma_l}\sqrt{\frac{\Gamma(\frac{3}{q})}{\Gamma(\frac{1}{q})}}$, and $\gamma_r = \frac{1}{\sigma_r}\sqrt{\frac{\Gamma(\frac{3}{q})}{\Gamma(\frac{1}{q})}}$. In this model, μ' is the mode, σ_l^2 and σ_r^2 are the variances of the left and right side respectively, and q is the decay rate. It is noticed that if $\sigma_l^2 = \sigma_r^2$ then the pdf coincides with the GG distribution, hence it is symmetric [Lee and Nandi, 1999]. For the symmetric cases, $c = 2$ represents the Gaussian distribution while $c = 1$ represents the Laplace distribution. If $\sigma_l^2 \neq \sigma_r^2$ then the pdf represents an asymmetric model. As shown in Fig. 3, the AGG family of distributions can represent wide range of unimodal probability density functions by changing the shape parameter q.

Suppose that we are given a data set of observations of scalar values $\mathbf{x} = \{x_1, \ldots, x_N\}$. Then, the log likelihood function of the AGG pdf is given by

$$\begin{aligned} \ln L_{\text{agg}} = N \ln\left(\frac{q\gamma_a}{\Gamma(\frac{1}{q})}\right) &- \sum_{i=1, x_i<\mu'}^{N} \gamma_l^q(\mu' - x_i)^q \\ &- \sum_{i=1, x_i\geq\mu'}^{N} \gamma_r^q(x_i - \mu')^q. \end{aligned} \tag{9}$$

We can optimize the parameters in the AGG model so as to maximize Eq. (9). For details, see 'Fitting to an AGG distribution (S3.2)' in Supplementary Materials.

3.3 Scoring scheme

Given the model for describing the probability density of the WNNSO values of main orthologs M_{wnnso}^+ and the model for describing the probability density of the WNNSO values of paralogs M_{wnnso}^-, the *synteny score* of a homologous gene pair h_m whose WNNSO value is x is defined by

$$\text{Syn\_Score}(h_m) = \ln \frac{P(x|M_{\text{wnnso}}^+)}{P(x|M_{\text{wnnso}}^-)}. \tag{10}$$

As shown in Supplementary Fig. S4, the score function given by Eq. (10) is not monotonically increasing function of x, although the score function is desired to be motononically increasing function because it is considered that the homologous gene pairs which have greater WNNSO value are more likely to be main orthologs. Thus, we modify Eq. (10) so that the score function be monotonically increasing. The modified score function is given by

$$\begin{aligned} &\text{Modified\_Syn\_Score}(h_m) \\ &= \begin{cases} \ln \frac{P(x|M_{\text{wnnso}}^+)}{P(x|M_{\text{wnnso}}^-)} & \text{for } x \leq \hat{x} \\ \\ \ln \frac{P(\hat{x}|M_{\text{wnnso}}^+)}{P(\hat{x}|M_{\text{wnnso}}^-)} + \delta(x - \hat{x}) & \text{for } x > \hat{x}, \end{cases} \end{aligned} \tag{11}$$

where $\hat{x}$ is the WNNSO value at which the score function given by Eq. (10) takes the maximum value, and δ is a extremely small value. Supplemenrary Fig. S4 demonstrates that the modified score function given by Eq. (11) is a monotonically increasing function of x. Analogously, the modified version of the *sequence score* Modified_Seq_Score(h_m) is defined by Supplemental Eq. (S15). For details, see 'Sequence score (S4.1)' in Supplementary Materials.

OASYS integrates the information about the extent of the gene order conservation and the extent of the protein sequence conservation by taking the weighted sum of the modified synteny score given by Eq. (11) and the modified sequence score given by Supplemental Eq. (S15). The integrated score is given by

$$\begin{aligned}&\text{Integrated\_Score}(h_m)\\ &= w_{\text{syn}}\text{Modified\_Syn\_Score}(h_m) + w_{\text{seq}}\text{Modified\_Seq\_Score}(h_m),\end{aligned} \tag{12}$$

where w_{syn} and w_{seq} denote the weight for the modified synteny score and the modified sequence score, respectively. The OASYS program has the weight ratio option, which can specify the weight ratio $\frac{w_{\text{syn}}}{w_{\text{seq}}}$. The default value for the weight ratio is set at 1.0. The effect of the weight ratio parameter is described in 'Effect of $\frac{w_{\text{syn}}}{w_{\text{seq}}}$ parameter (S4.3)' in Supplementary Materials.

References

[Akaike, 1974] Akaike,H. (1974) A new look at statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716-723.

[Alexeyenko *et al.*, 2006] Alexeyenko,A., Lindberg,J., Perez-Bercoff,A. Sonnhammer,E.L.L. (2006) Overview and comparison of ortholog databases, *Drug Discovery Today: Technol.*, **3**, 137-143.

[Altschul *et al.*, 1990] Altschul,S.F., Gish,W., Miller,W., Myers,E.W., Lipman,D.J. (1990) Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403-410.

[Bandyopadhyay *et al.*, 2006] Bandyopadhyay,S., Sharan,R., Ideker,T. (2006) Systematic identification of functional orthologs based on protein network comparison, *Genome Res*, **16**, 428-435.

[Benson *et al.*, 2009] Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Sayers,E.W. (2009) GenBank, *Nucleic Acids Res*, **37**, D26-31.

[Cannon and Young, 2003] Cannon,S.B., Young,N.D. (2003) OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies, *BMC Bioinformatics*, **4**, 35.

[Dewey *et al.*, 2006] Dewey,C.N., Huggins,P.M., Woods,K., Sturmfels,B., Pachter,L. (2006) Parametric alignment of Drosophila genomes, *PLoS Comput Biol*, **2**, e73.

[Fitch, 1970] Fitch,W.N. (1970) Distinguishing Homologous from Analogous Proteins, *Syst. Zool.*, **19**, 99-113.

[Fu *et al.*, 2007] Fu,Z., Chen,X., Vacic,V., Nan,P., Zhong,Y., Jiang, T. (2007) MSOUR: a high-throughput ortholog assignment system based on genome rearrangement, *J. Comput. Biol.*, **14**, 1160-1175.

[Hirsh and Fraser, 2001] Hirsh,A.E., Fraser,H.B. (2001) Protein dispensability and rate of evolution, *Nature*, **411**, 1046-1049.

[Hubbard *et al.*, 2009] Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K. *et al.* (2009) Ensembl 2009, *Nucleic Acids Res*, **37**, D690-697.

[Jordan *et al.*, 2002] Jordan,I.K., Rogozin,I.B., Wolf,Y.I., Koonin,E.V. (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria, *Genome Res*, **12**, 962-968.

[Lee and Nandi, 1999] Lee,J.Y., Nandi,A.K. (1999) Maximum likelihood parameter estimation of the asymmetric generalised gaussian family of distributions, *Proc. SPW-HOS.*

[Li *et al.*, 2003] Li,L., Stoeckert,C.J., Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Res*, **13**, 2178-2189.

[Miller and Thomas, 1972] Miller,J.H., Thomas,J.B. (1972) Detectors for discrete-time signals in non-Gaussian noise, *IEEE Transaction on Information Theory*, **18**, 241-250.

[Remm *et al.*, 2001] Remm, M. and Storm, C.E. and Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, *J. Mol. Biol.*, **314**, 1041-1052.

[Rivera *et al.*, 1998] Rivera,M.C., Jain,R., Moore,J.E., Lake,J.A. (1998) Genomic evidence for two functionally distinct gene classes, *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 6239-6244.

[Sayers *et al.*, 2009] Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K. *et al.* (2009) Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res*, **37**, D5-15.

[Singh *et al.*, 2008] Singh,R., Xu,J., Berger,B. (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection, *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 12763-12768.

[Sjölander, 2004] Sjölander,K. (2004) Phylogenomic inference of protein molecular function: advances and challenges, *Bioinformatics*, **20**, 170-179.

[Tatusov *et al.*, 1997] Tatusov,R.L., Koonin,E.V., Lipman,D.J. (1997) A genomic perspective on protein families, *Science*, **278**, 631-637.

[Tatusov *et al.*, 2003] Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., *et al.* (2003) The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4**, 41.

[Tesei and Regazzoni, 1998] Tesei,A., Regazzoni,C.S. (1998) HOS-based generalized noise pdf models for signal detection optimization, *Signal Processing*, **65**, 267-281.

[Vandepoele *et al.*, 2002] Vandepoele,K., Saeys,Y., Simillion,C., Raes,J., Van De Peer,Y. (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice, *Genome Res*, **12**, 1792-1801.

[Vilella *et al.*, 2009] Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R., Birney,E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates, *Genome Res*, **19**, 327-335.

[Zheng *et al.*, 2005] Zheng,X.H., Lu,F., Wang,Z.Y., Zhong,F., Hoover,J., *et al.* (2005) Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs, *Bioinformatics*, **21**, 703-710.