Cube-DB: Detection of Functional Divergence in Human Protein Families - Supplementary Material

Zong Hong Zhang, Kavitha Bharatham, Sharon Chee, and Ivana Mihalek *

Bioinformatics Institute 30 Biopolis Street, #07-01 Matrix, Singapore 138671

---

version = 2.19 of fff.tex 2005 Feb 5
running title: Cube-DB

Calculating conservation and specialization scores in Cube-DB

In this Supplement we list explicitly the algebraic expression evaluated by the application behind the Cube-DB database. For details, see (1)

Premliminaries

Let the distribution of residue types acceptable at position i, in the group of orhologous sequences s be given be given by $F_s = (f_{s1}, \ldots, f_{s20})$, where f_{sa} stands for the frequency of occurence of type a at position i, in the group s. In the average case, this distribution is expected (therefore the superscript x) to evolve as

$$F_s^{(x)}(t) = P(t)F_s(0) = P(t) \begin{pmatrix} f_{s1}(0) \\ \vdots \\ f_{s20}(0) \end{pmatrix}. \quad (1)$$

$P_{ba}(t)$ is the probability of the amino acid type indexed by a mutating to the one indexed by b in time t. The matrix P, in turn, is generated by the rate matrix A (2),

$$P(t) = e^{At}, \quad (2)$$

A used in Cube-DB is from Veerassamy *et al.* (3).

Conservation

Observed conservation (superscript o) is evaluated as $c^{(o)} = 1 - S^{(o)}$, with

* Corresponding author.
Email addresses: zhangzh@bii.a-star.edu.sg (Zong Hong Zhang), kavithab@bii.a-star.edu.sg (Kavitha Bharatham), sharonc@bii.a-star.edu.sg (Sharon Chee), ivanam@bii.a-star.edu.sg (Ivana Mihalek).

$$S^{(o)} = -\sum_a f_a log f_a. \tag{3}$$

To include the exchangeability of types we use the difference of observed entropy from its expected valuei in the average case:

$$S^{(m)} = S^{(o)} - S^{(x)} = -\sum_a f_a log f_a + \sum_a f_a^{(x)}(t_{_{eff}}) log f_a^{(x)}(t_{_{eff}}), \tag{4}$$

where f_a stands for the frequency observed in the alignment, and $f_a^{(x)}(t_{_{eff}})$ for the expected frequency of the type a in time $t_{_{eff}}$

The effective time $t_{_{eff}}$ is estimated by comparing all positions in the alignment (1).

Overlap in the amino acid types

In Cube-DB, overlap in the amino-acid types in paralogous groups s and t is evaluated as

$$o_{st}^{(o)} = \sum_{a=1}^{20} f_{sa} f_{ta}, \text{ such that } \sum_a f_a^2 = 1, \tag{5}$$

where index o again stands for the observed value, and f_{sa}, f_{ta} are the frequencies of residue type a in protein groups s and t respectively.

Including similarity:

$$o_{st}^{(m)}(t) = o_{st}^{(o)} - o_{st}^{(x)} = o_{st}^{(o)} - F_s^{(x)}(t)^T F_t^{(x)}(t) \tag{6}$$

where T indicates transpose, and $F_s^{(x)}(t)^T F_t^{(x)}(t)$ is the size of the overlap we expect in the average case.

Discriminants

The specialization score for positions behaving according to discrimanant model (conserved within each group of orthologues from different species, different in paralogues from the same specie) is evaluated as

$$dis^{(l)} = \sum_{g_1} \left((1 - c_{g_1}) + \sum_{g_2} o_{g_1 g_2} \right). \tag{7}$$

Determinants

The specialization score for positions behaving according to determinant model (conserved in the reference group, different but of arbitrary degree of conservation in non-reference groups)

is evaluated as:

$$det^{(l)} = (1 - c_t) + \sum_g o_{tg}. \tag{8}$$

References

[1] Bharatham, K., Zhang, Z., and Mihalek, I. (2011) Determinants, discriminants, conserved residues - a heuristic approach to detection of functional divergence in protein families. *PLoS One,* p. 10.1371/journal.pone.0024382.

[2] Felsenstein, J. Inferring Phylogenies chapter 13 Sinauer Associates, Sunderland, Mass. (2004).

[3] Veerassamy, S., Smith, A., and Tillier, E. (2003) A transition probability model for amino acid substitutions from blocks. *Journal of Computational Biology,* **10**(6), 997–1010.