A Simple Sensitive Program for Detecting Internal Repeats in Sets of Multiply Aligned Homologous Proteins

Yufeng Zhai, and Milton H. Saier, Jr.*

Division of Biology, University of California at San Diego, La Jolla, CA 92093-0116

Abstract

We designed a simple but sensitive program, IntraCompare, for identifying internal repeats in families of homologous proteins. The protein sequences are aligned (Clustal X), the regions to be compared are selected, and all potential repeat sequences are compared with all others. The output provides comparison scores (GAP program) expressed in standard deviations.

Introduction

Proteins and their associated functions evolve as a consequence of inherited genetic alterations. Thus the huge spectrum of proteins found in living organisms has it roots in genetic events operating on a set of ancestral genes. Gene duplication is one of the major events responsible for increased complexity in the protein world (Ohno, 1970). Duplication within a gene often gives rise to non-overlapping regions of the protein sequence that share sequence similarity. These repeats vary considerably in length and degree of sequence similarity, depending on the nature of the duplication, the time interval elapsed since the duplication occurred and the rate of sequence divergence (Saier, 1994). The polyglutamine tracts of the Huntington's disease gene product, huntingtin (Davies and Ramsden, 2001), and the large repetitions containing multiple domains found in the cytoskeletal protein titin (Gerull *et al.*, 2002) provide interesting examples.

Over the past decade, our lab and others have been concerned with molecular archaeological studies aimed at revealing the origins and evolutionary histories of permeases (Saier, 1999). The importance of repeats to an understanding of biological function resides not only in their frequencies among known sequences, but also in their abilities to confer multiple binding sites and complementary functions. Internal gene duplication can also provide evidence concerning the evolution pathway taken for the appearance of a family of transport proteins (Saier and Tseng, 1999). Demonstration of dissimilar repeat elements in different families can be used to establish independent origin (Saier, 2001). They also

*For correspondence. Email msaier@ucsd.edu; Tel. (858) 534-4084; Fax. (858) 534-7108.

play an important role in the appearance of novel transporter types proposed to follow the pathway: channel proteins → secondary active transporters → primary active transporters (Saier, 2000a).

Methods used to identify internal proteinaceous repeats exhibiting high sequence similarities are relatively straightforward. Detecting homologous repeats when similarities are low, however, represents a considerably more challenging endeavor. In this communication, we provide a program that facilitates such endeavors.

There are currently several tools available for detecting repeat sequences in proteins. For example, the REP program (Andrade *et al.*, 2000; http://www. emblheidelberg.de/~andrade/paper/rep/search.html] and the SMART program (Schultz *et al.*, 2000) can be used for this purpose. The REP program is a homology-based method which uses an iterative algorithm to estimate the significance of possible repeats in a sequence. For some purposes, this method is more sensitive and selective than conventional homology search procedures. The internal repeat identification program implemented in the SMART program is called Prospero (Mott, 2000). This program compares protein sequences using the Smith-Waterman algorithm and assesses statistical significance using a novel accurate formula.

The traditional PSI-BLAST program (Altschul *et al.*, 1997) and the HMMER program (Eddy, 1998) may facilitate repeat detection. Repeats are usually indicted by (1) a region in the query sequence that aligns with two distinct regions in the second protein, and (2) different regions of the query sequence that align with a homologous region in a second protein.

These methods have been widely used, but often they are unable to correctly predict the duplication because of weak sequence similarity. Here we describe a novel program (IntraCompare) with enhanced sensitivity because it identifies internal duplications in members of an entire protein family rather than in a single protein. Since this method is based on all possible comparisons for potentially homologous regions in members of large families, the probability of identifying internal duplications is greatly enhanced.

Description

The IntraCompare program is implemented with the GAP program (Devereaux *et al.*, 1984) which can be used to compare any two protein sequences. Based on our experience, homology between two proteins is considered proven if comparable regions of these

Further Reading

- **MALDI-TOF Mass Spectrometry in Microbiology**
  Edited by: M Kostrzewa, S Schubert (2016)
  www.caister.com/malditof

- *Aspergillus* and *Penicillium* in the Post-genomic Era
  Edited by: RP Vries, IB Gelber, MR Andersen (2016)
  www.caister.com/aspergillus2

- **The Bacteriocins: Current Knowledge and Future Prospects**
  Edited by: RL Dorit, SM Roy, MA Riley (2016)
  www.caister.com/bacteriocins

- **Omics in Plant Disease Resistance**
  Edited by: V Bhadauria (2016)
  www.caister.com/opdr

- **Acidophiles: Life in Extremely Acidic Environments**
  Edited by: R Quatrini, DB Johnson (2016)
  www.caister.com/acidophiles

- **Climate Change and Microbial Ecology: Current Research and Future Trends**
  Edited by: J Marxsen (2016)
  www.caister.com/climate

- **Biofilms in Bioremediation: Current Research and Emerging Technologies**
  Edited by: G Lear (2016)
  www.caister.com/biorem

- **Microalgae: Current Research and Applications**
  Edited by: MN Tsaloglou (2016)
  www.caister.com/microalgae

- **Gas Plasma Sterilization in Microbiology: Theory, Applications, Pitfalls and New Perspectives**
  Edited by: H Shintani, A Sakudo (2016)
  www.caister.com/gasplasma

- **Virus Evolution: Current Research and Future Directions**
  Edited by: SC Weaver, M Denison, M Roossinck, et al. (2016)
  www.caister.com/virusevol

- **Arboviruses: Molecular Biology, Evolution and Control**
  Edited by: N Vasilakis, DJ Gubler (2016)
  www.caister.com/arbo

- *Shigella*: **Molecular and Cellular Biology**
  Edited by: WD Picking, WL Picking (2016)
  www.caister.com/shigella

- **Aquatic Biofilms: Ecology, Water Quality and Wastewater Treatment**
  Edited by: AM Romaní, H Guasch, MD Balaguer (2016)
  www.caister.com/aquaticbiofilms

- **Alphaviruses: Current Biology**
  Edited by: S Mahalingam, L Herrero, B Herring (2016)
  www.caister.com/alpha

- **Thermophilic Microorganisms**
  Edited by: F Li (2015)
  www.caister.com/thermophile
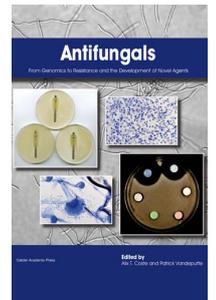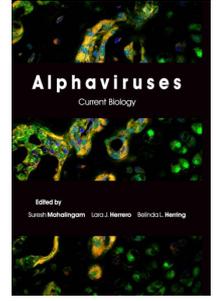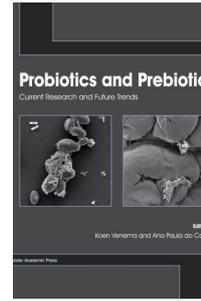
- **Flow Cytometry in Microbiology: Technology and Applications**
  Edited by: MG Wilkinson (2015)
  www.caister.com/flow

- **Probiotics and Prebiotics: Current Research and Future Trends**
  Edited by: K Venema, AP Carmo (2015)
  www.caister.com/probiotics

- **Epigenetics: Current Research and Emerging Trends**
  Edited by: BP Chadwick (2015)
  www.caister.com/epigenetics2015

- *Corynebacterium glutamicum*: **From Systems Biology to Biotechnological Applications**
  Edited by: A Burkovski (2015)
  www.caister.com/cory2

- **Advanced Vaccine Research Methods for the Decade of Vaccines**
  Edited by: F Bagnoli, R Rappuoli (2015)
  www.caister.com/vaccines

- **Antifungals: From Genomics to Resistance and the Development of Novel Agents**
  Edited by: AT Coste, P Vandeputte (2015)
  www.caister.com/antifungals

- **Bacteria-Plant Interactions: Advanced Research and Future Trends**
  Edited by: J Murillo, BA Vinatzer, RW Jackson, et al. (2015)
  www.caister.com/bacteria-plant

- *Aeromonas*
  Edited by: J Graf (2015)
  www.caister.com/aeromonas

- **Antibiotics: Current Innovations and Future Trends**
  Edited by: S Sánchez, AL Demain (2015)
  www.caister.com/antibiotics

- *Leishmania*: **Current Biology and Control**
  Edited by: S Adak, R Datta (2015)
  www.caister.com/leish2

- *Acanthamoeba*: **Biology and Pathogenesis (2nd edition)**
  Author: NA Khan (2015)
  www.caister.com/acanthamoeba2

- **Microarrays: Current Technology, Innovations and Applications**
  Edited by: Z He (2014)
  www.caister.com/microarrays2

- **Metagenomics of the Microbial Nitrogen Cycle: Theory, Methods and Applications**
  Edited by: D Marco (2014)
  www.caister.com/n2
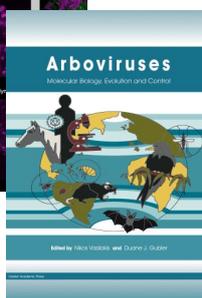
Table 1. Members of the PHS family used in the analysis.

Abbreviation	Database description	Organism	Size	GI #
PHO84 Sce	Inorganic phosphate transporter pho84	*Saccharomyces cerevisiae*	587	1346710
GvPT Gve	Phosphate transport protein	*Glomus versiforme*	521	2147899
Pho-5+ Ncr	Repressible high-affinity phosphate permease	*Neurospora crassa*	569	536860
AtPT1 Ath	Phosphate transporter	*Arabidopsis thaliana*	524	1502428
AtPT2 Ath	Phosphate transporter	*Arabidopsis thaliana*	534	15224985
PT1 Stu	Inorganic phosphate transporter	*Solanum tuberosum*	540	5053118

proteins (60 residues or more) exhibit a comparison score in excess of 9 standard deviations. This value corresponds to a probability of less than 10^{-19} that the observed degree of similarity occurred by chance (Dayhoff *et al.*, 1983). We have designed IntraCompare to start with a multiple sequence alignment in Clustal X format (Thompson *et al.*, 1997) since Clustal X is one of the most popular and reliable sequence alignment tools currently available.

In the command line of the IntraCompare program, the user may define the region of potential internal duplication as, for example, can be determined using the AveHAS program (Zhai and Saier, 2001a) or the WHAT program (Zhai and Saier, 2001b), both developed in our laboratory. AveHAS is a CGI program that can use the multiple sequence alignment generated with the Clustal X program as its input to calculate average hydropathy, average amphipathicity and average similarity as a function of alignment position. These qualities are of particular value for the characterization of membrane proteins. WHAT provides hydropathy and amphipathicity values as a function of

residue number plus additional information for individual proteins. IntraCompare systematically compares the regions defined, and standard deviation values are outputted in tab delimited format which can be easily loaded into any spreadsheet program such as Excel for further data analysis.

Application to the PHS Family

The Phosphate:H$^+$ Symporter (PHS) (TC #2.A.1.9) family within the Major Facilitator Superfamily (MFS) is presumed to be of ancient origin, and therefore one would expect it to be ubiquitous in the major domains of life (Pao *et al.*, 1998; Saier, 2000b). Two well-characterized members of the PHS family are the Pho84 inorganic phosphate transporter of *Saccharomyces cerevisiae* (Bun-Ya *et al.*, 1991) and the GvPT phosphate transporter of *Glomus versiforme* (Harrison and van Buuren, 1995).

We chose 6 members of the PHS family (Table 1) to make a multiple sequence alignment, and this sequence alignment was analyzed by the AveHAS
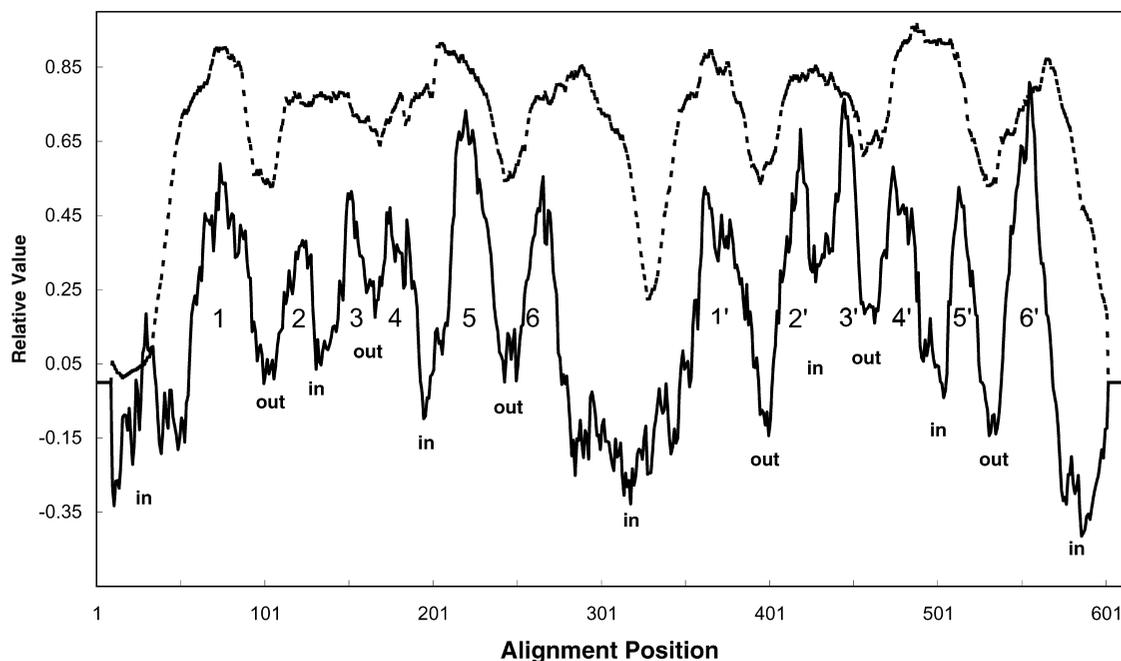

Figure 1. Average hydropathy (solid line) and average similarity (dashed line) for the multiple alignment of the 6 members of the PHS family presented in Table 1. The AveHAS program was used to generate the plots. The 12 TMSs are labeled 1-6 (first repeat) and 1′-6′ (second repeat). in, in the cytoplasm; out, in the extracytoplasmic medium.

program. Figure 1 shows the average hydropathy and similarity plots. There are 8 regions of conservation revealed by the output of AveHAS, and the symmetry of the plot suggests that the 2 halves might have arisen from a single primordial sequence. Thus, each half contains 6 hydrophobic peaks corresponding to 6 putative transmembrane α-helices, and the graphic display of the first half of the plot resembles the second half. Since the N-and C-termini of all characterized MFS permeases are cytoplasmic, the similarity plot reveals that the inter-TMS cytoplasmic loops are better conserved than the extracytoplasmic loops. The possibility that an internal duplication event occurred during evolution of this family has been documented previously (see Pao *et al.*, 1998 and references cited therein).

We analyzed each of the 6 sequences in the PHS family using the REP and SMART programs, but neither of these programs found the duplication. According to the AveHAS results, we split the sequence alignment into two halves, corresponding to alignment positions 1 to 328 and 329 to 610. IntraCompare compared the first halves with the second halves, i.e., 36 comparisons. This number of comparisons increases exponentially with the number of sequences in the alignment. Two comparisons made by IntraCompare clearly revealed the internal duplication and established homology. One is the first half of AtPT2 Ath compared with the second half of Pho-5+ Ncr, and the other is the first half of GvPT Gve compared with the second half of Pho84 Sce. The standard deviations of these comparisons are 9.1 and 11.5, respectively, clearly establishing homology (Saier, 1994). From the IntraCompare results, we conclude that proteins of the PHS family, and therefore probably all members of the MFS, arose by an internal gene duplication event (Pao *et al.*, 1998).

Conclusion

We have described a novel program (IntraCompare), written in PERL language, which uses Clustal X sequence alignments as input to calculate the similarity of regions in the alignment which may be internally duplicated.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Andrade, M.A., Ponting, C., Gibson, T., and Bork, P. 2000. Homology-based method for identification of protein repeats using statistical significance estimates. J. Mol. Biol. 298: 521–537.

Bun-Ya, M., Nishimura, M., Harashima, S., and Oshima, Y. 1991. The PHO84 gene of *Saccharomyces cerevisiae* encodes an inorganic phosphate transporter. Mol. Cell Biol. 11: 3229–3238.

Davies, S., and Ramsden, D.B. 2001. Huntington's disease. Mol. Pathol. 54: 409–413.

Dayhoff, M.O., Barker, W.C., and Hunt, L.T. 1983. Establishing homologies in protein sequences. Methods Enzymol. 91: 524–545.

Devereux, J., Haeberli, P., and Smithies, O. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12: 387–395.

Eddy, S.R. 1998. Profile hidden Markov models. Bioinformatics 14: 755–763.

Gerull, B., Gramlich, M., Atherton, J., McNabb, M., Trombitas, K., Sasse-Klaassen, S., Seidman, J.G., Seidman, C., Granzier, H., Labeit, S., Frenneaux, M., and Thierfelder, L. 2002. Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. Nat. Genet. (in press).

Harrison, M.J., and van Buuren, M.L. 1995. A phosphate transporter from the mycorrhizal fungus *Glomus versiforme*. Nature 378: 626–629.

Mott, R. 2000. Accurate formula for P-values of gapped local sequence and profile alignments. J. Mol. Biol. 300: 649–659.

Ohno, S. 1970. *Evolution by Gene Duplication.* Springer-Verlag, Heidelberg, Germany.

Pao, S.S., Paulsen, I.T., and Saier, M.H., Jr. 1998. Major facilitator superfamily. Microbiol. Mol. Biol. Rev. 62: 1–34.

Saier, M.H., Jr. 1994. Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. Microbiol. Revs. 58: 71–93.

Saier, M.H., Jr. 1999. Genome archeology leading to the characterization and classification of transport proteins. Curr. Opin. Microbiol. 2: 555–561.

Saier, M.H., Jr. 2000a. Vectorial metabolism and the evolution of transport systems. J. Bacteriol. 182: 5029–5035.

Saier, M.H., Jr. 2000b. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol. Mol. Biol. Rev. 64: 354–411.

Saier, M.H., Jr. 2001. Evolution of transport proteins. In *Genetic Engineering. Principles and Methods*, Vol. 23 (J.K. Setlow, ed.). New York: Kluwer Academic/Plenum Press, pp. 1–10.

Saier, M.H., Jr., and Tseng, T.-T. 1999. Evolutionary origins of transmembrane transport systems. In *Transport of Molecules Across Microbial Membranes*, Symposium 58, Society for General Microbiology (J.K. Broome-Smith, S. Baumberg, C.J. Stirling and F.B. Ward, eds.), Cambridge University Press, Cambridge, UK, pp. 252–274.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., and Bork, P. 2000. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. 28: 231–234.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24: 4876–4882.

Zhai, Y., and Saier, M.H., Jr. 2001a. A web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of multiply aligned homologous proteins. J. Mol. Microbiol. Biotechnol. 3: 285–286.

Zhai, Y., and Saier, M.H., Jr. 2001b. A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. J. Mol. Microbiol. Biotechnol. 3: 501–502.