Drosophila Genomic Sequence Annotation Using the BLOCKS+ Database

Jorja G. Henikoff[1] and Steven Henikoff[1,2]

[2]Howard Hughes Medical Institute, [1]Fred Hutchinson Cancer Research Center, Seattle, Washington 98109-1024 USA

A simple and general homology-based method for gene finding was applied to the 2.9-Mb *Drosophila melanogaster Adh* region, the target sequence of the Genome Annotation Assessment Project (GASP). Each strand of the entire sequence was used as query of the BLOCKS+ database of conserved regions of proteins. This led to functional assignments for more than one-third of the genes and two-thirds of the transposons. Considering the enormous size of the query, the fact that only two false-positive matches were reported emphasizes the high selectivity of protein family-based methods for gene finding. We used the search results to improve BLOCKS+ by identifying compositionally biased blocks. Our results confirm that protein family databases can be used effectively in automated sequence annotation efforts.

Sequence similarity searches for detecting protein relationships have become so popular that one method for doing this is now familiarly described by the verb "to blast." Detecting a hit in a sequence data bank is frequently the best clue as to the function of a gene, so that sequence similarity searching is de rigeur for any genomic annotation effort. A routine annotation strategy is to first arrive at a gene model (Fields and Soderlund 1990), translate it into protein, then use the predicted protein as query of sequence data banks (Pearson and Lipman 1988; Altschul et al. 1990). Most entrants in the GASP (Genome Annotation Assessment Project) study attempted to find accurate gene models, and their success in doing this is the basis for assessment of their performance (Reese et al. 2000a). Some methods used sequence similarity searches of cDNA databases to aid in predicting accurate gene models. Another method (GeneWise) screened gene models against a protein family database. Our method differs in that we dispensed entirely with the gene modeling step, using the full genomic segment to query a protein family database. The rationale is that protein sequence is so rich in information that even this simple approach will be sufficiently sensitive to find the genes and assign functions to them.

Our method is nearly a decade old. Using protein queries to search DNA databases translated in all six frames, which was introduced 12 years ago (Henikoff and Wallace 1988; Pearson and Lipman 1988), has since become a standard procedure, especially for searching EST databases (Adams et al. 1991). Alternatively, a DNA query can be translated for searching protein sequence or protein family databases, such as the BLOCKS database (Henikoff and Henikoff 1991). Entries in the BLOCKS database are ungapped multiple alignments of conserved regions of proteins, averaging four BLOCKS per protein family. In a search, detection of multiple BLOCKS representing a family are combined into a hit. In a translated search, BLOCKS are combined into a hit even when they are in different frames on the same strand. Earlier, we reported the detection of a *Pseudomonas cepacia* regulatory gene (*dgdR*) and protein family homology for *dgdA* within a 4-kb genomic segment used as query (Henikoff and Henikoff 1991); both had been missed because of frameshift sequencing errors. This example emphasized the fact that translated searching allows for gene detection and family assignment without requiring assumptions as to the presence of ORFs or the accuracy and completeness of the sequence used as query.

With the release of the first complete chromosome sequence, *Saccharomyces cerevisiae* chromosome III (Oliver et al. 1992), we applied this fully automated method to a >300-kb genomic segment (Henikoff and Henikoff 1994). Each frame of the entire sequence was used to search a 1992 version of the BLOCKS database, and the results for each strand were combined to make gene predictions. We found 37 significant hits, of which 34 were genes discovered by others, 1 was a new gene not detected by others, and 2 were judged to be false positives. This number of hits represented only 40% of what could be found using pairwise approaches, an expected result considering the low coverage of the 1992 BLOCKS database relative to what was available in sequence data banks. When we repeated the search on a 1993 version of the BLOCKS database, 10 more genes were found, a consequence of expansion of the BLOCKS database from 504 to 619 protein families (Henikoff and Henikoff 1994).

At the time of the GASP study (June 1999), the BLOCKS database had increased to >2000 protein families. Most of the increase is due to supplementation of the original BLOCKS database, which is based on fami-

[1]Corresponding author.
E-MAIL steveh@fhcrc.org; FAX (206) 667-5889.

10:543–546 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

lies catalogued in PROSITE (Hofmann et al. 1999), with protein families documented in other compendiums: PRINTS (Attwood et al. 1999), PFAM (Bateman et al. 1999), ProDom (Corpet et al. 1999), and DOMO (Gracy and Argos 1998). This "BLOCKS+" database (Henikoff et al. 1999) was made nonredundant by applying the blocks-vs.-blocks LAMA searching method (Pietrokovski 1996) to eliminate protein families that shared significant sequence similarities. The substantial increase in coverage, coupled with the value of protein family-based annotation, encouraged us to try out this simple method on a highly complex genomic sequence, and GASP provided an opportunity.

RESULTS

A total of 109 hits were submitted to GASP. Of these, 93 proved to be within 78 of the 222 protein-coding genes annotated by Ashburner et al. (1999). Another 13 hits corresponded to 12 of 17 annotated transposons. Although we could not determine whether all of the 78 genes were correctly annotated, all of the transposon hits are recognizable as such because the blocks represent various families of retrotransposon-encoded proteins, such as reverse transcriptases and aspartyl proteases.

Of the remaining three hits, two are undoubtedly false positives. Both are single block hits that we ought to have removed from the results list during the final manual scrutiny. A hit to Block BL01253G was marginally detectable ($E = 0.33$); it is the only alignment found for a family that is represented by eight blocks and, so, is highly questionable. The other was a hit at $E = 0.01$ to BP02591, which we noticed in hindsight is a compositionally biased (cysteine-rich) block. It is possible that more careful scrutiny or more refined criteria for accepting hits would have avoided these false-positive errors.

The remaining hit, to the M2 peptidase family (PF01401), was also predicted as a gene by several other GASP participants, including GeneWise, which also characterized it as an M2 peptidase. This hit lies within the 2-kb region that separates two genes annotated by Ashburner et al. (1999) as "*Ance*" and "*Acyp*." It is interesting that *Ance* encodes an M2 peptidase as does the gene beyond *Acyp*, "*DS00180.5*," and so we predicted a cluster of three M2 peptidase genes in the region. It appears that there is a partial duplication of *Ance*, perhaps part of the *Ance* transcription unit that would be alternatively processed.

DISCUSSION

What Went Right?

The results of GASP indicate that excellent specificity can be achieved using a simple general approach that does not depend in any way on gene modeling. Although we used manual scrutiny as a final step, the criteria we used for acceptance could be refined and implemented in software, which would make our method fully automated. Our overall performance was roughly comparable to that of GeneWise, which used gene models rather than six-frame translations to search a protein family database (Birney and Durbin 2000). This suggests that six-frame translated searching of genomic sequence is adequately sensitive for multi-megabase queries. Our detection level was somewhat higher than that of GeneWise, most likely because we searched a more comprehensive family database rather than because of any important methodological difference.

As protein family databases expand, homology-based methods should become increasingly valuable for annotating complex genomes. The very low level of false positive GASP predictions by the two protein family methods (BLOCKS+ and GeneWise) is encouraging. Hits to protein family databases provide immediate clues as to the function of a gene. For instance, all 12 transposons that we detected were identifiable as such based on the protein families hit. Hits to blocks also pinpoint the location of conserved motifs, which is important for further functional characterization. The recent InterPro initiative (http://www.ebi.ac.uk/interpro/) promises to make protein family databases even more useful and accessible, and we anticipate future integration of BLOCKS+ with InterPro.

Any improvement of the BLOCKS+ database increases its value for all types of searching. The GASP exercise provided us with a list of compositionally biased blocks that can be disregarded or removed when performing large-scale annotations. We had previously used algorithmic criteria to identify compositionally biased blocks; however, we have not found these criteria to be satisfactory. We prefer the empirical approach used here. Thus, the GASP exercise has directly helped to improve our system.

What Went Wrong?

Although the hits we predicted included very few wrong genes (high specificity), many genes were missed (low sensitivity). This could be due to the absence of protein families in the BLOCKS+ database or to the stringent statistical cutoffs used to overcome the high background in such a large search. We predicted a very low percentage of coding region bases, as is to be expected for any method that only predicts conserved regions of proteins.

Translated searching gets noisier as coding regions become diluted by noncoding regions, diminishing the advantage of our approach over methods that use gene models. As the quality of gene models improves, they miss fewer coding regions, and so the high background that our approach encounters becomes less tolerable. Furthermore, the trend toward large-scale sequencing centers with higher quality standards means that there are fewer frameshift and other sequencing errors that can cause gene modeling to fail. Because the simplicity, generality, and automation of translated searching comes at a high cost, we do not expect that our method will be widely adopted over methods based on gene models. Only 3 of the 222 GASP genes were not predicted by any of the participating protein coding region prediction programs, and these 3 genes may be cDNA cloning artifacts (Table 2 in Reese et al. 2000b). BLOCKS+ did not predict these three genes either, so all genes predicted by BLOCKS+ were predicted as protein coding regions by at least one program. We are impressed by the ability of current gene modeling programs to predict enough of a gene to make useful queries of sequence and protein family databases (Reese et al. 2000a), and it seems likely to us that this approach will ultimately prevail for most effective functional annotation. Protein family databases, such as those represented in BLOCKS+ and InterPro, are especially well-suited for effective annotation of gene models. They assist biologists in understanding protein function by providing a more complete view of homology and domain information than do sequence databases. We anticipate that continued development of protein family databases and tools will greatly improve the state of the art in functional annotation of genomes.

METHODS

We used the BLIMPS searching and BLKSORT postprocessing programs that are implemented for public use at our website (http://blocks.fhcrc.org) and e-mail (blocks@blocks.fhcrc.org) servers (Henikoff et al. 2000). These servers limit the size of queries, so users interested in performing searches on the scale of GASP may do so by installing a Unix version of the BLIMPS system (Henikoff et al. 1995) (ftp://ncbi.nlm.nih.gov/repository/blocks/unix/blimps). To augment coverage of BLOCKS+ for GASP, we also searched blocks derived from the SMART 3.0 database (Schultz et al. 2000) for a total of 2430 protein families represented by 10,637 blocks.

Each strand of the 2.9-Mb query sequence was searched against this augumented BLOCKS+ database (31 million alignments per strand), and hits with expected (*E*) values better than 10 were sorted by location for each strand. It was immediately noticed that a small number of blocks accounted for a disproportionate number of high scoring alignments. These blocks were judged to be compositionally biased and were removed from the database, and the search was rerun. Each search required a few processor days (on a SUN Sparcstation 20) and postprocessing required a few hours. Given the large size of the query and the high background of chance hits, it was not feasible to save twilight zone alignments for assembly into multiple block hits, and so we did not expect chance hits involving more than one block to be reported. However, single block hits required a stricter standard, and they were arbitrarily removed from the results list if they scored worse than *E* = 1. At this point, single block hits, overlapping hits, and questionable multiple block hits were individually examined by the authors for plausibility. Each block responsible for a single block hit was examined for compositional bias, and if bias was noted, the hit was removed from the results list.

ACKNOWLEDGMENTS

We thank the GASP team for motivating this study. This work was supported by a grant from NIH.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McCombie, and J.C. Venter. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Ashburner, M., S. Misra, J. Roote, S.E. Lewis, R. Blazej, T. Davis, C. Doyle, R. Galle, R. George, N. Harris, G. Hartzell, D. Harvey, L. Hong, and K. Houston. 1999. An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the *Adh* region. *Genetics* **153:** 179–219.

Attwood, T.K., D.R. Flower, A.P. Lewis, J.E. Mabey, S.R. Morgan, P. Scordis, J.N. Selley, and W. Wright. 1999. PRINTS prepares for the new millenium. *Nucleic Acids Res.* **27:** 220–225.

Bateman, A., E. Birney, R. Durbin, S.R. Eddy, R.D. Finn, and E.L.L. Sonnhammer. 1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27:** 260–262.

Birney, E. and R. Durbin. 2000. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* (this issue).

Corpet, F., J. Gouzy, and D. Kahn. 1999. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res.* **27:** 263–267.

Fields, C.A. and C.A. Soderlund. 1990. gm: A practical tool for automating DNA sequence analysis. *Comp. Appl. Biosci.* **6:** 263–270.

Gracy, J. and P. Argos. 1998. Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics* **14:** 164–173.

Henikoff, J.G., E.A. Greene, S. Pietrokovski, and S. Henikoff. 2000. Increased coverage of protein families with the Blocks Database servers. *Nucleic Acids Res.* **28:** 228–230.

Henikoff, S. and J.G. Henikoff. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* **19:** 6565–6572.

———. 1994. A protein family classification method for analysis of large DNA sequences. In *Proceedings of the 27th Annual Hawaii International Conference on Systems Sciences,* pp. 265–274. Institute of Electrical and Electronics Engineers, New York, NY.

Henikoff, S. and J.C. Wallace. 1988. Detection of protein similarities using nucleotide sequence databases. *Nucleic Acids Res.* **16:** 6191–6204.

Henikoff, S., J.G. Henikoff, W.J. Alford, and S. Pietrokovski. 1995. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene* **163:** GC17–GC26.

Henikoff, S., J.G. Henikoff, and S. Pietrokovski. 1999. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15:** 471–479.

Hofmann, K., P. Bucher, L. Falquet, and A. Bairoch. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27:** 215–219.

Oliver, S.G., Q.J.M. van der Aart, M.L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, and J.P.G. Ballesta. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357:** 38–46.

Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Pietrokovski, S. 1996. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.* **24:** 3836–3845.

Reese, M.G., N. Harris, G. Hartzell, U. Ohler, and S. Lewis. 2000a. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* (this issue).

Reese, M.G., D. Kulp, H. Tammana, and D. Haussler. 2000b. Genie—Gene finding in *Drosophila melanogaster*. *Genome Res.* (this issue).

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

Received February 9, 2000; accepted February 28, 2000.