Biochemical and Biophysical Research Communications **273,** 90–94 (2000)
doi:10.1006/bbrc.2000.2910, available online at http://www.idealibrary.com on IDEAL®

Characterization of Five Novel Human Genes in the 11q13-q22 Region

Kevin P. O'Brien,* Isabel Tapia-Páez,† Mona Ståhle-Bäckdahl,*
Darek Kedra,† and Jan P. Dumanski†
**Department of Medicine and †Department of Molecular Medicine, Karolinska Hospital, S-171 76, Stockholm, Sweden*

Received May 22, 2000

The redundancy of sequences in dbEST has approached a level where contiguous cDNA sequences of genes can be assembled, without the need to physically handle the clones from which the ESTs are derived. This is termed EST based *in silico* gene cloning. With the availability of sequence chromatogram files for a subset of ESTs, the quality of EST sequences can be ascertained accurately and used in contig assembly. In this report, we performed a study using this approach and isolated five novel human genes, *C11orf1-C11orf5*, in the 11q13-q22 region. The full open reading frames of these genes were determined by comparison with their orthologs, of which four mouse orthologs were isolated (*c11orf1, c11orf2, c11orf3* and *c11orf5*). These genes were then analyzed using several proteomics tools. Both C11orf1 and C11orf2 are nuclear proteins with no other distinguishing features. C11orf3 is a cytoplasmic protein containing an ATP/GTP binding site, a signal peptide located in the N-terminus and a similarity to the *C. elegans* protein "Probable ARP 2/3 complex 20kD subunit." C11orf4 is a peptide which displays four putative transmembrane domains and is predicted to have a cytoplasmic localization. It contains signal peptides at the N- and C-termini. C11orf5 is a putative nuclear protein displaying a central coiled coil domain. Here, we propose that this purely EST-based cloning approach can be used by modestly sized laboratories to rapidly and accurately characterize and map a significant number of human genes without the need of further sequencing.

The Human Genome Project (HGP) is a large international collaboration which intends to resolve the nucleotide sequence of the entire human genome and catalog all genes encoded therein (1). Already in 1999, a milestone in this undertaking was achieved, with the complete sequence of the first human chromosome, that of chromosome 22, being reported (2). Since only 2–5% of the human genome accounts for coding sequences, *de novo* gene identification is notoriously difficult. The accuracy of gene prediction programs varies dramatically and suffers from underprediction and overprediction, therefore output should be treated as a rough guide as opposed to a true transcription map. The advent of the Expressed Sequence Tags (EST) project has proved to be a excellent supplemental tool in gene identification (3). To date, over 2 million ESTs have been released in the EST database (dbEST) and this number has continued to rise. It is likely that over 80% of all human genes have at least one corresponding EST in dbEST (4).

In silico cloning, the electronic cloning of a gene without the need to physically handle DNA, is fast becoming the method of choice for gene identification. This can be performed by the clustering of ESTs together to form contiguous sequences, from which possible open reading frames (ORFs) can be predicted (5). This approach, combined with the mapping of EST clusters to human-hamster radiation hybrids lead to the formation of the first comprehensive human gene map, with over 30,000 genes being positioned (6, 7). However, there are two main problems with using ESTs: (i) ESTs are based on single sequence reads and are therefore prone to sequence errors and (ii) ESTs are derived from cDNA libraries and therefore may contain artifacts such as introns, repeats/ALUs or erroneous chimeric transcripts. Hence, automatically assembled clusters of ESTs can be comprised of both expressed and unexpressed sequences and should be treated with caution. Fortunately, the sequence chromatogram files of a subset of ESTs are publicly available and can be used in semiautomatic assembly of EST clusters in which bad sequence can be corrected. Artifacts derived from cDNA library synthesis, e.g. unspliced introns, can be identified by comparison with

Accession Numbers: AJ250229, AJ249980, AJ250344, AJ250392, AJ250393, AJ250230, AJ249981, AJ250356, and AJ250394.

0006-291X/00 $35.00

either the mouse ortholog or with other normal EST sequences. This is possible due to the high redundancy of dbEST (5).

We used this method of cloning in the identification of five novel human genes in the 11q13-q22 region, *C11orf1-C11orf5,* and characterize their ORFs. We also employ this method to isolate the mouse orthologs of these genes (four complete ORFs, one partial). We make use the comparison of human/mouse sequences to identify the true translation initiation and termination sites in these novel genes.

MATERIALS AND METHODS

Chromosome 11q mapping information and positions of EST clusters were obtained from the NCBI/NIH "Gene Map" website (www.ncbi.nlm.nih.gov/genemap) (6). The *Blast* family of programs was used for database searches on the NCBI/NIH servers (www.ncbi.nlm.nih.gov/blast). Sequence chromatogram files for ESTs were imported by ftp (file transfer protocol) from genome.wustl.edu and assembled using the *asp* program (8), the *phrap2gap* (9) and the *gap4* programs of the Staden package (10). Open reading frames and amino acid sequences were predicted using the *translate* tool (www.expasy.ch/tools/dna.html). Predicted proteins were aligned using the *clustal_x* program (11). Pairwise identity and similarity predictions were calculated using *clustal_x* and *boxshade* programs (11). Dot matrix plots were produced using the *dotter* program (12). Following assembly and identification of ORFs, putative proteins were analyzed using the following tools: the *blast* family for homology searches; *scanprosite* (www.expasy.ch/prosite), *pfam* (pfam.wustl.edu), *blocks* (www.blocks.fhcrc.org) and *smart* (smart.embl-heidelberg.de) to search for known protein families and domains; *coils* (www.ch.embnet.org), *paircoil* (nightingale.lcs.mit.edu/cgi-bin/score) and *multicoil* (nightingale.lcs.mit.edu/cgi-bin/multicoil) to detect coiled-coiled domains; and *psort* (psort.nibb.ac.jp:8800) in the prediction of protein sorting signals and localization sites.

RESULTS

Using the NCBI/NIH website Gene map '98 (6), chromosome 11q13-q22 was examined for the presence of STSs which had corresponding EST clusters. Those clusters which did not represent previously characterized genes were analyzed further for quantity of matching mouse and human ESTs in the databases. When a cluster was deemed to have sufficient representation (>25 ESTs), the clone names of all matching ESTs were obtained by *Blastn* search. The clones for which sequence chromatogram files were likely to be available (i.e., clone names starting with "a," "y" and "z" for human, "m," "u" and "v" for mouse) were imported and assembled. The resulting contigs were then analyzed for the presence of ORFs, a process which included the removal of sequence chromatogram files which were considered to contain cDNA synthesis artifacts. This lead to the identification of the ORFs of five novel human genes in the 11q13-q22 region; *C11orf1, C11orf2, C11orf3, C11orf4,* and *C11orf5*. The mouse orthologs of four of these genes were also isolated; *c11orf1, c11orf2, c11orf3,* and *c11orf5*. The mouse ortholog of *c11orf4* was represented in the database by only one EST, which contained only the 5′ end of the ORF. In most cases dot matrix plot comparisons between the human and mouse orthologs clearly delineated the position of the ORFs, since the 5′ and 3′ untranslated regions show lack of conservation (Fig. 1). Where it was unclear whether an ATG codon was in fact the true first methionine, i.e., human/mouse conservation seen upstream of the ATG and no Kozak consensus sequence (ACC(A/G)CCATGG) present, then alternative strategies were employed. The presence of either stop codons upstream of a methionine in one or more of the orthologs and the conservation of an ATG in more than one other organism was considered as proof of a translation initiation site. A summary of the details of genes cloned and the features of their putative proteins are shown (Tables 1 and 2, respectively).

The presence of homologous ESTs from other organisms in dbEST was performed using the *tblastx* search option to obtain a rough indication of gene evolution. *C11orf1* was represented by only human and mouse ESTs, while *C11orf2* was represented also by rat ESTs and *C11orf4* by chicken ESTs. *C11orf3* was shown to be present in, rat, rabbit, zebrafish, *C. elegans,* tomato and *Arabidopsis. C11orf5* was the gene most represented in the databases with ESTs from over 20 organisms e.g. rat, rabbit, zebrafish, drosophila, *C. elegans,* tomato and *Arabidopsis.* It should be noted that dbEST is dominated by human and mouse ESTs and lack of ESTs for selected organisms does not necessarily indicate absence of a gene. However, *tblastx* results can be used to broadly outline the evolutionary conservation of a gene.

DISCUSSION

In silico gene cloning is often seen as a supplemental approach to the more conventional laboratory gene cloning. In the past gene prediction programs tended to heavily overestimate the number of genes/exons in a sequence and laborious screening was required to eliminate false positives. The advent of ESTs revolutionized gene identification and it is now possible to rapidly and accurately detect the vast majority of transcribed regions/genes in a genomic sequence (1, 5, 6). However, it is not always necessary to have the complete genomic sequences from a locus in order to positionally clone a gene *in silico,* to identify its protein or to suggest a function for a gene. The sequence redundancy of dbEST has approached a level where contiguous cDNA sequences of genes can be assembled, without the need to physically handle the clones from which the ESTs are derived. With the availability of sequence chromatogram files for a subset of ESTs, the quality and accuracy of EST sequences can be determined unequivocally, for example our *C11orf5* contig (839 bp) is composed of 56 quality sequence chromatogram files.

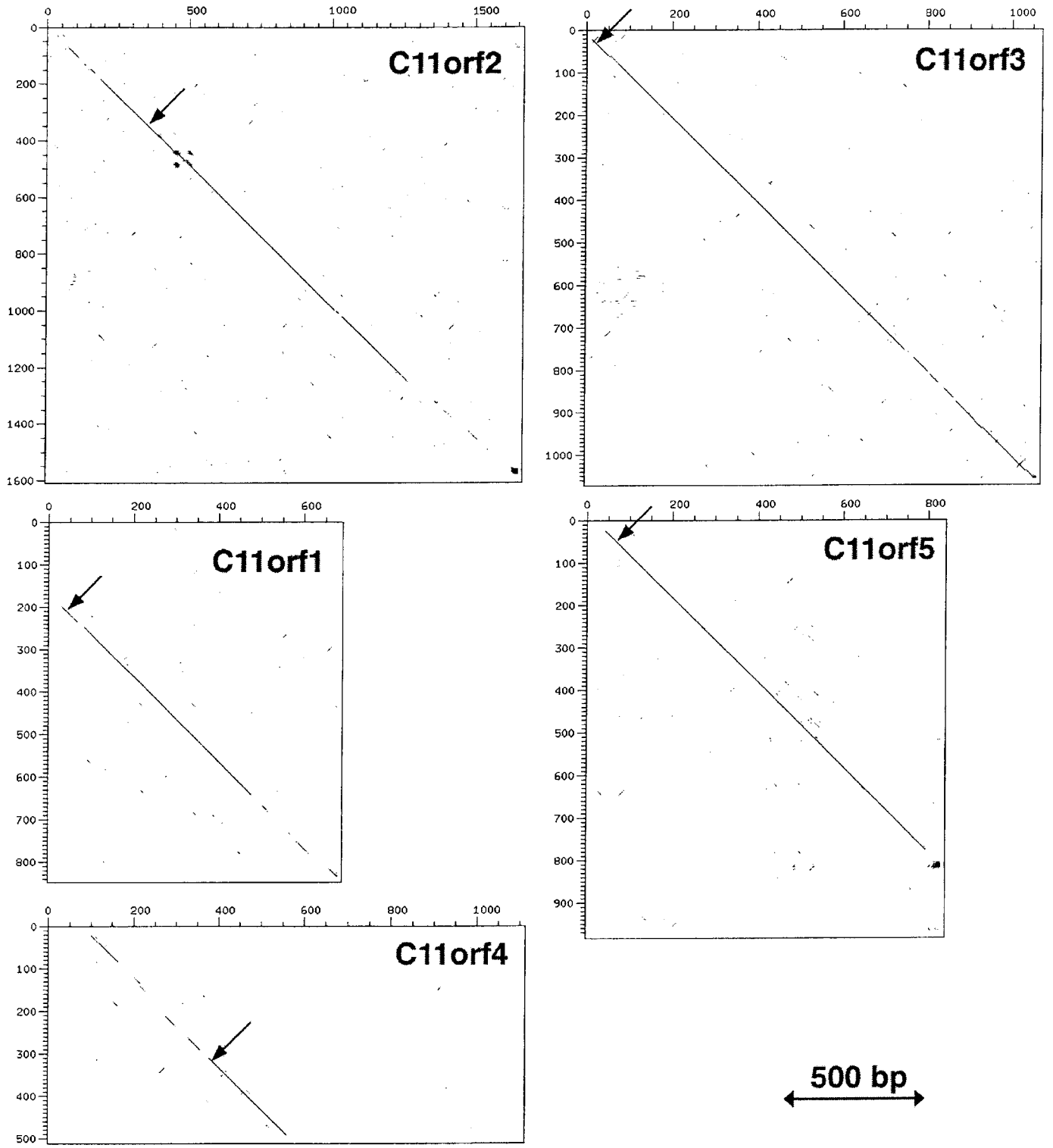

FIG. 1. Dot matrix plots comparing the nucleotide sequence of human genes with their mouse orthologs. *C11orf4* is plotted against a single mouse EST covering the 5′ end of its mouse ortholog. Plots were performed using the *dotter* program. Human mRNA sequences are placed on the X-axis, mouse on the Y-axis. Arrows indicate the position of the first methionines. A generally decreased homology and stop codons in one or both orthologs were found upstream of these methionines indicating that they were true translation initiation sites. Drawn to scale.

TABLE 1

Gene	Accession No.	Species	Length	No. Readings	STS	Map	PolyA
C11orf1	AJ250229	H. sapiens	688	11	SHGC-33774	11q21-22	Yes
C11orf2	AJ249980	H. sapiens	1663	20	WI-12265	11q22	Yes
C11orf3	AJ250344	H. sapiens	1072	29	SHGC-64737	11q13	Yes
C11orf4	AJ250392	H. sapiens	1110	19	WI-17324	11q13	No
C11orf5	AJ250393	H. sapiens	839	56	D11S1757	11q21-22	Yes
c11orf1	AJ250230	M. musculus	844	4	N/A	N/A	Yes
c11orf2	AJ249981	M. musculus	1601	16	N/A	N/A	Yes
c11orf3	AJ250356	M. musculus	1066	26	N/A	N/A	Yes
c11orf5	AJ250394	M. musculus	977	25	N/A	N/A	Yes

Note. Human and mouse genes cloned. "Accession No." indicates the accession number of the newly cloned genes. "Length" indicates sizes of sequence contigs obtained in base pairs. "No. Readings" denotes the number of sequence chromatogram files comprising the contig. "STS" indicates the name of the STS encompassed by a cloned gene and "map" indicates its subchromosomal position. "PolyA" shows whether a contig contained a PolyA tail. "N/A" indicates information not available.

Hence EST based cloning is roughly analogous to shotgun cloning sequencing in which randomly generated subclones are end-sequenced and assembled to form a contiguous sequence. The difference with the EST based approach is that, first, the sequencing is not performed by the assembling investigator, and second, the clones are not truly random, tending instead to contain the polyA tail in a random sized clone. Therefore, purely EST based *in silico* cloning is most successful in medium to small sized genes in which the open reading frame is close to the polyA tail. The limitation with this method is that the true 5′ end of a gene is not always possible to ascertain and can only be truly determined with a combination of Rapid Amplification of cDNA Ends (5′ RACE), full length cDNA sequencing or Northern blot analysis. If a polyA tail is seen in more than one EST (to rule out false priming in cDNA synthesis) and a corresponding polyadenylation signal is present, the 3′ end of a gene can be easily pinpointed. However, in the case of *C11orf4* no EST containing a polyA tail could be found and therefore the true 3′ end is not known. Nevertheless, using the data presently available the strength of this approach lies in its ability to isolate ORFs.

While planning chromosome walking experiments during the mapping the constitutional translocation t(11;22)(q23;q11) (Tapia-Páez *et al.*, manuscript in preparation) we noted that chromosome 11 contained a large number of STS (Sequence Tagged Sites) and a comprehensive map (13). We therefore chose the region centromeric to the breakpoint to study using the above described method in combination with mapped STSs. We isolated and characterized five human genes, *C11orf1-C11orf5*, in the region 11q13-q22 (Table 1). The full ORFs of these genes were ascertained via comparison with their orthologs, of which we cloned four orthologs in mouse (*c11orf1*, *c11orf2*, *c11orf3*, and *c11orf5*). Using the various proteomics programs available (see Materials and Methods), we identified various features which gives clues as to cellular localization

TABLE 2

Protein	Human length (aa)	Mouse ortholog length (aa)	Similarity	Identity	Predicted features	Similarity to other proteins
C11orf1	150	166	73%	79%	Nuclear protein	None
C11orf2	304	300	91%	91%	Two regions of low complexity in N-terminus Nuclear protein	None
C11orf3	242	241	98%	96%	ATP/GTP Binding site Signal peptide aa 1-23 Cytoplasmic protein	ACC007070 (57%, *Arabidopsis thaliana*) Q18491 (57%, *C. elegans*)
C11orf4	210	N/A	N/A	N/A	Four transmembrane domains N- and C-terminal signal peptides Cytoplasmic protein	4507545 (74%, *H. sapiens*)
C11orf5	229	229	100%	99%	Coiled coil domain aa 123-165 Nuclear protein	Z93388 (51%, *C. elegans*) CAA21276 (67%, *S. pombe*) AAF00144 (68%, *O. sativa*)

Note. Putative peptides of cloned genes. "Length" indicates length in amino acids of putative peptides in human and mouse. "Similarity" and "Identity" denote the similarity and identity between human and mouse orthologs. "Homologous Proteins" indicates known and putative proteins showing >50% similarity upon *Blastp* analysis. "N/A" indicates information not available.

and function of the putative peptides of these five novel genes (Table 2). Both C11orf1 and C11orf2 are likely to be nuclear proteins with no other distinguishing features. C11orf3 is thought to be a cytoplasmic protein containing an ATP/GTP binding site and a signal peptide located in the N-terminus. Upon *blastp* analysis we detected two homologous proteins; a *C. elegans* protein (Probable ARP 2/3 complex 20 kD subunit, Accession No. Q18491) (14) and an *Arabidopsis* protein with an unknown function (Accession No. AC007070). C11orf4 is a peptide which displays four putative transmembrane domains and is predicted to have a cytoplasmic localization. It contains signal peptides on both N- and C-termini. It displays a homology to the human protein transmembrane 7 superfamily member 1 (Accession No. 4507545). Finally, C11orf5 is a putative nuclear protein displaying a central coiled coil domain, displaying homology to *C. elegans*, *S. pombe*, and *Oryza sativa* proteins with unknown functions (Accession Nos. Z93388, CAA21276 and AAF00144, respectively). Coiled-coil domains have been found in many proteins and can form stable, rod-like structures that mediate protein–protein interactions via formation of helices coiled around each other (15, 16).

In conclusion, one of the goals of the Human Genome Project is the development of cDNA resources, to supplement the annotating of genomic sequence, which is considered to be urgently needed (1). We believe that the approach used in this study can be rapidly applied to the accurate characterization of a notable portion of human genes. Genes of small to medium sizes can be identified and cloned without the need of further sequencing, whereas larger genes can also be characterized with additional small scale sequencing of EST clones. It is conceivable for a modestly sized laboratory to isolate, characterize and assign a chromosomal location to a large number of genes. It is possible to concentrate gene cloning in a region of biological interest and provide a scaffold for positional cloning projects or to perform a genome wide scan. All this is feasible using data already available in public databases.

ACKNOWLEDGMENTS

This work was supported by grants from the Swedish Cancer Foundation, the Swedish Medical Research Council, the Cancer Society in Stockholm, the Berth von Kantzow Fond, the Ake Wiberg's Foundation, the Karolinska Hospital and the Karolinska Institutet to J.P.D.

REFERENCES

1. Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. (1998) *Science* **282,** 682–689.
2. Dunham, I., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., Ainscough, R., Almeida, J. P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K. N., Beasley, O., Bird, C. P., Blakey, S., Bridgeman, A. M., Buck, D., and Burgess, J., *et al.* (1999) *Nature,* in press.
3. Boguski, M. S. (1995) *Trends Biochem. Sci.* **20,** 295–296.
4. Strachan, T., and Read, A. P. (1996) Human Molecular Genetics, pp. 335–365, Wiley & Sons, New York, NY.
5. Banfi, S., Guffanti, A., and Borsani, G. (1998) *Trends Genet.* **14,** 80–81.
6. Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T. C., McKusick, K. B., Beckmann, J. S., Bentolila, S., Bihoreau, M., Birren, B. B., Browne, J., Butler, A., Castle, A. B., Chiannilkulchai, N., Clee, C., Day, P. J., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., Bentley, D. R., *et al.* (1998) *Science* **282,** 744–746.
7. Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B. B., Butler, A., Castle, A. B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., and Hudson, T. J., *et al.* (1996) *Science* **274,** 540–546.
8. Wendl, M. C., Dear, S., Hodgson, D., and Hillier, L. (1998) *Genome Res.* **8,** 975–984.
9. Dear, S., Durbin, R., Hillier, L., Marth, G., Thierry-Mieg, J., and Mott, R. (1998) *Genome Res.* **8,** 260–267.
10. Staden, R. (1994) *in* Methods in Molecular Biology (Griffin, A. M., and Griffin, H. G., Eds.), Vol. 25, pp. 9–170, Humana Press Inc, Totawa, NJ.
11. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) *Nucleic Acids Res.* **25,** 4876–4882.
12. Sonnhammer, E. L., and Durbin, R. (1995) *Gene* **167,** 1–10.
13. Perlin, M. W., Duggan, D. J., Davis, K., Farr, J. E., Findler, R. B., Higgins, M. J., Nowak, N. J., Evans, G. A., Qin, S., and Zhang, J. (1995) *Genomics* **28,** 315–327.
14. Welch, M. D., DePace, A. H., Verma, S., Iwamatsu, A., and Mitchison, T. J. (1997) *J. Cell Biol.* **138,** 375–384.
15. Wolf, E., Kim, P. S., and Berger, B. (1997) *Protein Sci.* **6,** 1179–1189.
16. Lupas, A., Van Dyke, M., and Stock, J. (1991) *Science* **252,** 1162–1164.