Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W649–W653***
DOI: 10.1093/nar/gkh455

ACMES: fast multiple-genome searches for short repeat sequences with concurrent cross-species information retrieval

Jeff Reneker[1,2], Chi-Ren Shyu[1,2,*], Peiyu Zeng[5], Joseph C. Polacco[3] and Walter Gassmann[4]

[1]Department of Computer Science, [2]Department of Health Management and Informatics, [3]Department of Agriculture Biochemistry and [4]Department of Plant Pathology, University of Missouri, Columbia, MO 65211, USA and [5]Department of Plant Science, University of Rhode Island, Kingston, RI, 02881, USA

Received February 15, 2004; Revised April 15, 2004; Accepted April 23, 2004

ABSTRACT

We have developed a web server for the life sciences community to use to search for short repeats of DNA sequence of length between 3 and 10 000 bases within multiple species. This search employs a unique and fast hash function approach. Our system also applies information retrieval algorithms to discover knowledge of cross-species conservation of repeat sequences. Furthermore, we have incorporated a part of the Gene Ontology database into our information retrieval algorithms to broaden the coverage of the search. Our web server and tutorial can be found at http://acmes.rnet.missouri.edu.

INTRODUCTION

Important nucleotide sequences are often reused throughout nature. For instance, whole genomic array analysis in yeast has revealed 22 PHO-regulated genes. The promoter regions of all but one of these contain at least one of the two core Pho4p binding sites, CACGTG and CACGTT (1). Once binding sites such as these have been determined experimentally, an investigator may wish to search for new, previously unknown genes that may be under the control of the same factors. Also, in bacterial pathogens, it is proposed that microsatellites regulate expression of some virulence factors (2). Another example is a region of an exon that codes for a specific motif within a protein. An investigator studying a protein with an interesting feature may wish to find other genes for proteins that share the same feature. Searching genomic sequences for a subsequence the size of a binding site (or shorter) is becoming increasingly important. However, these types of search can easily produce thousands of hits, especially with shorter queries. Several algorithms have been developed recently to search genomic sequences looking for short repeated patterns (3–12). All of these algorithms require the query to be greater than a certain minimum length, and there is an inverse relationship between this minimum length and the size of the data structure produced. Therefore, searching for short queries can lead to computer main memory deficits as more sequences are added to the structures. Unfortunately, adding more sequences is often the next logical step because multiple-species searches can help lead to discoveries about evolutionarily conserved sequences.

Another concern to address for these searches to produce meaningful results is how to reduce the number of reported hits, which can be in the thousands. To help with this goal, incorporating a sequence's annotation data into the search appears to be a promising approach. In this paper, we present our Advanced Content Matching Engine for Sequences (ACMES). Our engine is able to find exact matches to query sequences of any length, although we currently limit this length to between 3 and 10 000 bases. Also, our engine can quickly and efficiently search multiple species without the main memory constraints present with other systems, as discussed in the next section.

DESCRIPTION AND APPLICATION

Our search engine employs a novel hash function to preprocess each sequence. This hash function first converts short sequences of DNA (termed 'words') into integers and then sorts them. The information from these sorted integers is recorded into an index file and a corresponding data file which are written to a hard drive for permanent storage. During retrieval, a user's query is converted into an integer which serves as a key into the index file. Owing to page limit for this web server special issue, the detailed algorithms for this hash function can be viewed from the ACMES website.

Our hash map data structures are preprocessed and stored on the hard drive. During retrieval, we read only pages containing the appropriate hash bin. This allows our main memory usage to remain low even when processing multiple species, because

*To whom correspondence should be addressed. Tel: +1 573 882 3842; Fax: +1 573 882 3813; Email: shyuc@missouri.edu

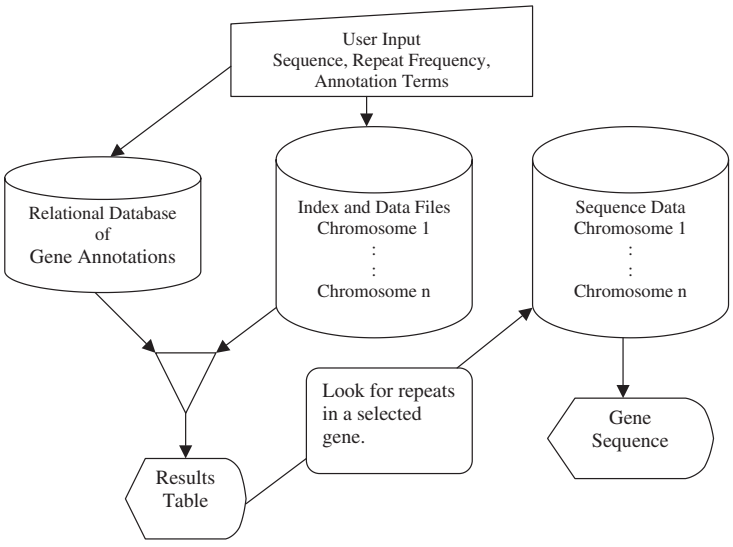
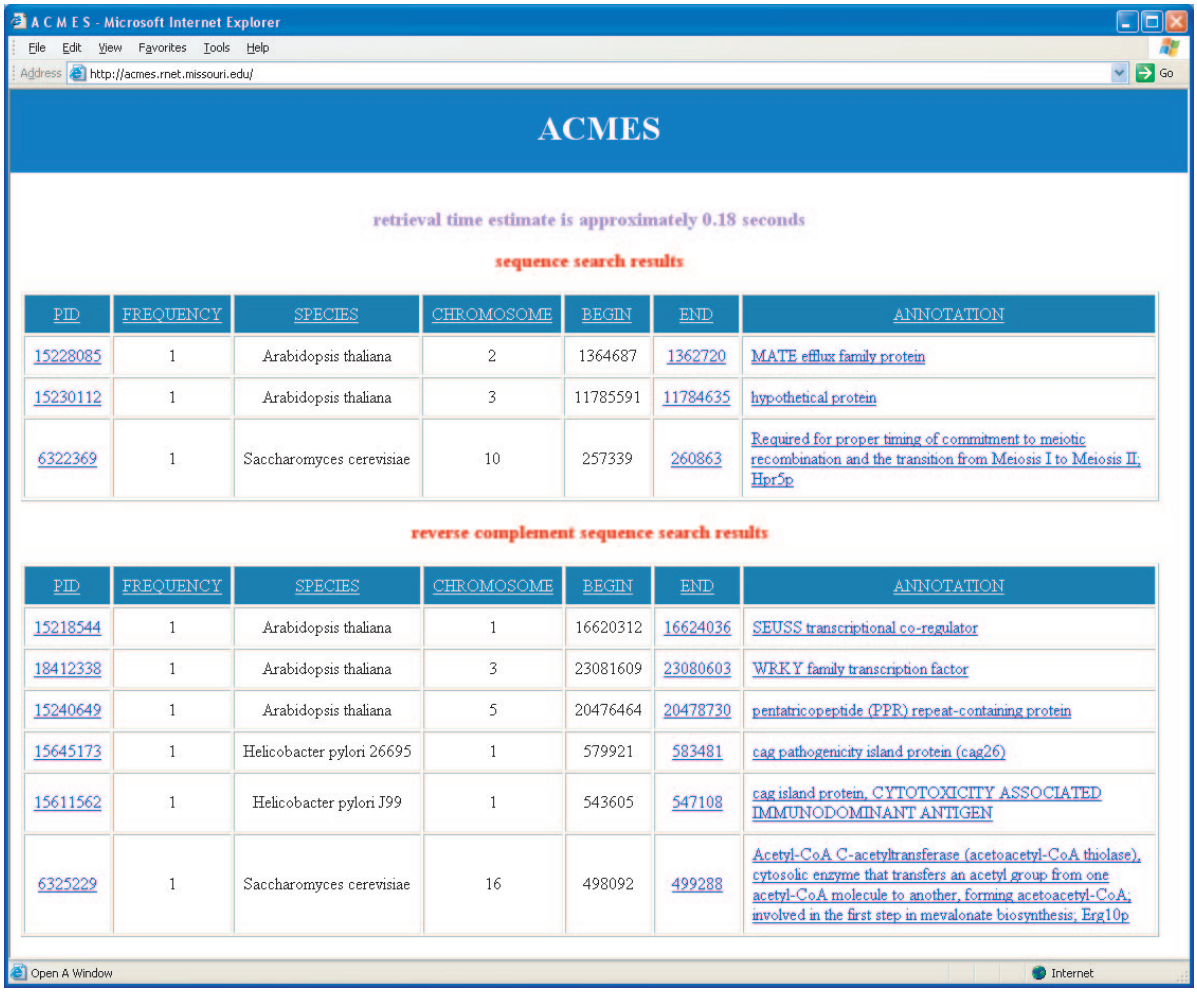


Figure 1. ACMES system diagram. The chromosome index file is searched based on the user's query. This index provides the location within the chromosome data file where a list of protein IDs can be found. Protein IDs must be present at least 'repeat frequency' times to be displayed in the results table. Also, protein IDs allow retrieval of annotation data, which is compared with the user's annotation terms. The results table contains all genes that map closest to the query sequence and contain at least one of the annotation terms, (if any are input). Users can then view the sequence of selected genes with the query sequence highlighted.



A C M E S - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://acmes.rnet.missouri.edu/ Go

ACMES

retrieval time estimate is approximately 0.18 seconds

sequence search results

PID	FREQUENCY	SPECIES	CHROMOSOME	BEGIN	END	ANNOTATION
15228085	1	Arabidopsis thaliana	2	1364687	1362720	MATE efflux family protein
15230112	1	Arabidopsis thaliana	3	11785591	11784635	hypothetical protein
6322369	1	Saccharomyces cerevisiae	10	257339	260863	Required for proper timing of commitment to meiotic recombination and the transition from Meiosis I to Meiosis II; Hpr5p

reverse complement sequence search results

PID	FREQUENCY	SPECIES	CHROMOSOME	BEGIN	END	ANNOTATION
15218544	1	Arabidopsis thaliana	1	16620312	16624036	SEUSS transcriptional co-regulator
18412338	1	Arabidopsis thaliana	3	23081609	23080603	WRKY family transcription factor
15240649	1	Arabidopsis thaliana	5	20476464	20478730	pentatricopeptide (PPR) repeat-containing protein
15645173	1	Helicobacter pylori 26695	1	579921	583481	cag pathogenicity island protein (cag26)
15611562	1	Helicobacter pylori J99	1	543605	547108	cag island protein, CYTOTOXICITY ASSOCIATED IMMUNODOMINANT ANTIGEN
6325229	1	Saccharomyces cerevisiae	16	498092	499288	Acetyl-CoA C-acetyltransferase (acetoacetyl-CoA thiolase), cytosolic enzyme that transfers an acetyl group from one acetyl-CoA molecule to another, forming acetoacetyl-CoA; involved in the first step in mevalonate biosynthesis; Erg10p

Open A Window Internet

Figure 2. Screen capture from the ACMES web site. All species were searched for 'ttaggctacctt' and its reverse complement. The protein ID, frequency, species, chromosome, beginning and ending locations and annotation are displayed. The protein ID contains a link to NCBI, the ending location contains a link to the gene sequence, and the annotation contains a link to re-search the database with the given terms plus closely related terms from Gene Ontology.

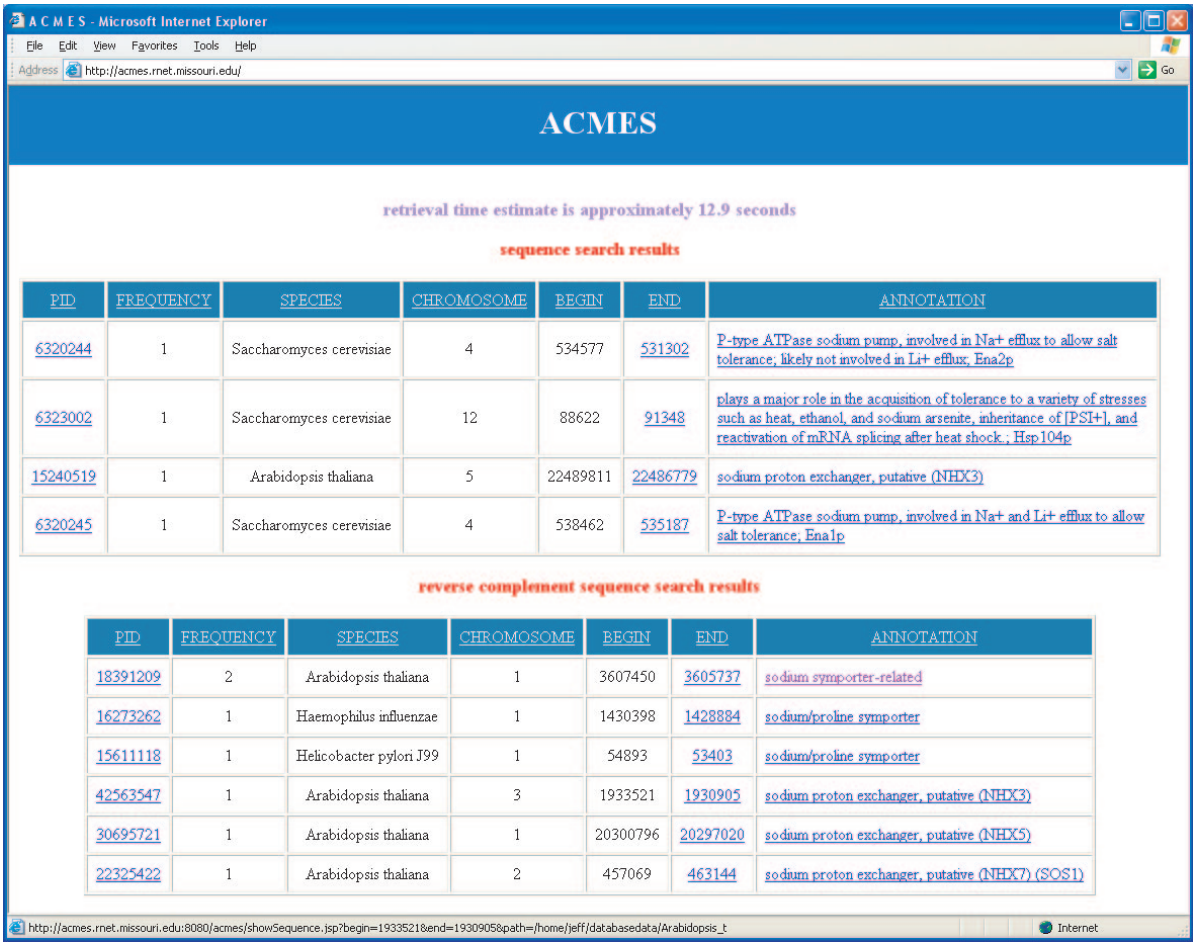


Figure 3. All species were searched for 'ttaggct' and its reverse complement. A total of 12 461 hits were retrieved. The database was re-searched by clicking on the annotation data 'sodium symporter-related' from one of the hits. The results shown are all genes that contain the original sequence and either 'sodium' or 'symporter-related' in their annotation data. The results are ranked according to the relevance of the annotation data.

earlier pages can be swapped for later pages after we finish using them. This approach also helps our IO costs to remain low, which allows our algorithm to perform quickly. For instance, our database presently consists of five different species: *Arabidopsis thaliana*, *Haemophilus influenzae*, *Helicobacter pylori* 26695, *Helicobacter pylori* J99 and *Saccharomyces cerevisiae*. We extracted 10 000 random sequences from *A. thaliana* chromosome 5 for each of 9 different lengths ranging from 4 to 1024 bases. For each query sequence, we searched each species in the database and recorded the retrieval times and the number of hits. Length 4 queries took 2.78 s and retrieved 751 536 hits on average. Length 8 queries took 0.072 s and retrieved 4993 hits on average. The query lengths continued to double up to 1024 bases. For length 1024 queries, the average retrieval time was 1.291 s and the average number of hits was 1.01. These tests were conducted on a dedicated server featuring dual Xeon IV 2.4 GHz processors, 2 GB RAM and a 120 GB EIDE 7200 r.p.m. hard disk running RedHat 9.0 RA.

Figure 1 shows the flow chart of the ACMES search engine. Current users of the site can find exact matches of query sequences as either disperse or tandem repeats. Disperse repeats are identical repeats separated by one or more bases, whereas tandem repeats are contiguous identical repeats. In addition, users can specify a minimum repeat frequency, which is the minimum number of times that a query must map to a gene in order to be displayed in the results table. This option can help reduce the number of reported hits for copious queries. Another, more focused approach to reduce the number of reported hits is also available. Users can enter one or several terms to describe the gene(s) that they are expecting in the results. After all of the hits have been retrieved from the database, the annotation data for each retrieved gene is searched for matching terms. Then, the user is presented with only those genes that contain both the query and at least one of the added terms. From the results tables, as shown in Figures 2 and 3, users can access additional genetic information at the National Center for Biotechnology Information (NCBI) at http://www.ncbi.nlm.nih.gov/. They can also view the genetic sequence, as shown in Figure 4, and/or re-search the database with new annotation terms.

Before each search for users' annotation terms, we add additional terms from the Gene Ontology (GO) database, http://www.geneontology.org. Each term is expanded with the most closely related terms available from GO. This is an attempt to include additional results that might otherwise have been missed. For instance, if a user entered 'programmed cell death', we would include the term 'apoptosis' in the

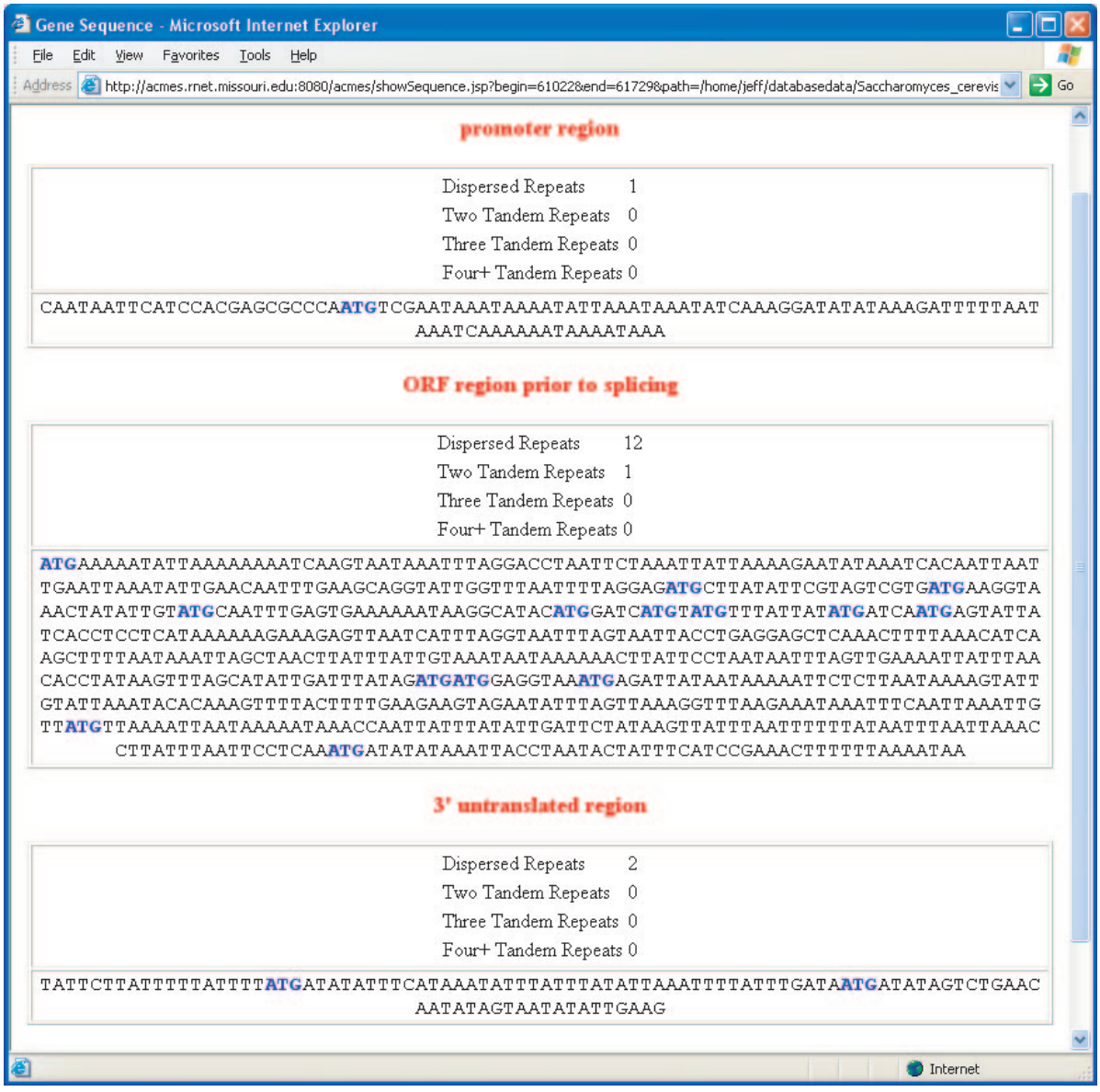


Figure 4. Screen capture of the sequence of the gene with protein ID 6226538 from the *Saccharomyces cerevisiae* mitochondria. The gene is divided into three regions: promoter, open reading frome prior to splicing and 3′-untranslated. The original query sequence ATG is highlighted on the web page and the numbers of disperse and tandem repeats are also displayed.

search. Also, during subsequent searches (begun by clicking on the annotation terms of an initial search result) retrieved results are ranked according to the product of term frequency and inverse document frequency of their annotation terms (13). Thus, results with less common terms are displayed first, while results with more common terms, such as 'putative protein', are displayed later.

CONTINUING WORK AND CONCLUSION

We continue to develop new algorithms concerning gene control and gene function. We are also in the process of adding the human genome to the database and will later add the mouse genome. As more sequences are added and as gene annotation for these sequences also improves, we anticipate richer, more meaningful searches for the life sciences community. We continue to incorporate more content from the Gene Ontology database into our algorithms. Therefore, the relevance of our retrievals will improve as this ontology continues to improve.

REFERENCES

1. Ogawa,N., DeRisi,J. and Brown,P. (2000) New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Mol. Biol. Cell.*, **11**, 4309–4321.
2. Hood,D.W., Deadman,M.E., Jennings,M.P., Bisercic,M., Fleischmann,R.D., Venter,J.C. and Moxon,R. (1996) DNA repeats identify novel virulence genes in *Haemophilus influenzae*. *Proc. Natl Acad. Sci. USA*, **93**, 11121–11125.

3. Benson,G. (1999) Tandem repeat finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
4. Adebiyi,E.F., Jiang,T. and Kaufmann,M. (2001) An efficient algorithm for finding short approximate non-tandem repeats. *Bioinformatics*, **17**, S5–S12.
5. Landau,G.M., Schmidt,J.P. and Sokol,D. (2001) An algorithm for approximate tandem repeats. *J. Comput. Biol.*, **8**, 1–18.
6. Castelo,A., Martins,W. and Gao,G. (2002) TROLL—tandem repeat occurrence locator. *Bioinformatics*, **18**, 634–636.
7. Kolpakov,R., Bana,G. and Kucherov,G. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.
8. Gusfield,D. (1997) *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge, UK.
9. Hauth,A. and Joseph,D. (2002) Beyond tandem repeats: complex pattern structures and distant regions of similarity. *Bioinformatics*, **18**, S31–S37.
10. 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.
11. Ning,Z., Cox,A.J. and Mullikin,J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.
12. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
13. Baeza-Yates,R. and Ribeiro-Neto,B. (1999) *Modern Information Retrieval*. ACM Press, New York, NY.