BMC Genomics

Database

Open Access

i-Genome: A database to summarize oligonucleotide data in genomes

Feng-Mao Lin[1], Hsien-Da Huang[2], Yu-Chung Chang[3] and Jorng-Tzong Horng*[1,4]

Address: [1]Department of Computer Science and Information Engineering National Central University, Chung-Li 320, Taiwan, [2]Department of Biological Science and Technology, Institute of Bioinformatics National Chiao Tung University, Hsin-Chu 300, Taiwan, [3]Department of Biotechnology, Ming Chuan University, Taipei 111, Taiwan and [4]Department of Life Science, National Central University, Chung-Li 320, Taiwan

Email: Feng-Mao Lin - meta@db.csie.ncu.edu.tw; Hsien-Da Huang - bryan@mail.nctu.edu.tw; Yu-Chung Chang - d80106@mailsrv3.ym.edu.tw; Jorng-Tzong Horng* - horng@db.csie.ncu.edu.tw

* Corresponding author

Published: 09 October 2004

BMC Genomics 2004, **5**:78 doi:10.1186/1471-2164-5-78

Received: 25 March 2004
Accepted: 09 October 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/78

© 2004 Lin et al; licensee BioMed Central Ltd.
This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Information on the occurrence of sequence features in genomes is crucial to comparative genomics, evolutionary analysis, the analyses of regulatory sequences and the quantitative evaluation of sequences. Computing the frequencies and the occurrences of a pattern in complete genomes is time-consuming.

Results: The proposed database provides information about sequence features generated by exhaustively computing the sequences of the complete genome. The repetitive elements in the eukaryotic genomes, such as LINEs, SINEs, Alu and LTR, are obtained from Repbase. The database supports various complete genomes including human, yeast, worm, and 128 microbial genomes.

Conclusions: This investigation presents and implements an efficiently computational approach to accumulate the occurrences of the oligonucleotides or patterns in complete genomes. A database is established to maintain the information of the sequence features, including the distributions of oligonucleotide, the gene distribution, the distribution of repetitive elements in genomes and the occurrences of the oligonucleotides. The database can provide more effective and efficient way to access the repetitive features in genomes.

Background

During the last decade, many genomes have been successfully and completely sequenced. Summarized information about the oligonucleotides in genomes provides biologists, who interests in the evolution and growth of genomes, to work in comparative genomics, oligonucleotide probe design, primer design and the analyses of genomic repetitive features.

The computation of the occurrences and the frequency of all oligonucleotides in a complete genome is very elaborate and time-consuming, especially when the genome size is very large, such as the human and mouse genomes.

Table 1: Applications and the relevant data in the database.

Database entries	Entry types	Biological applications
Oligonucleotide occurrences	Positions	1. Oligonucleotide analysis for regulatory sequences 2. Oligonucleotide probe design 3. Primer design
Oligonucleotide frequencies	Counts	1. Oligonucleotide analysis for regulatory sequences 2. Evolutionary analysis 3. Oligonucleotide probe design 4. Primer design
Gene coding regions	Positions	1. Oligonucleotide analysis for regulatory sequences 2. Evolutionary analysis 3. Oligonucleotide probe design 4. Primer design
Repetitive element frequencies (LINE, SINE, Alu, and so on)	Counts	Evolutionary analysis
Repetitive element occurrences	Positions	Evolutionary analysis
Tandem repeats	Positions	Prediction for genetic disease marker

A database that summarizes the occurrences and the frequencies of oligonucleotides in complete genomes can facilitate the biological and the statistical analyses of genomes.

The contents of the database can be used in many biological applications, such as comparative genomics and evolution analyses [1,2], the prediction of regulatory sequences by detecting the over-represented oligonucleotides [3-8] and primer/probe design based on the uniqueness of oligonucleotides [9]. Table 1 shows the biological applications of the database entries. The entries in the database are divided into two types, namely, the occurrence positions of oligonucleotides and the frequencies of oligonucleotides. The oligonucleotide occurrences and the oligonucleotide frequencies in both the coding regions and the non-coding regions are summarized. For instance, these information can be used in statistical analyses to study the over-representation of the regulatory sequences in upstream promoter regions in genes. van Helden *et al.* systematically searched the promoter regions of potentially co-regulated genes for over-represented oligonucleotides which may be transcription factor binding sites [3]. They presented a simple and fast method for isolating DNA binding sites for transcription factors from families of co-regulated genes, illustrating their results using *Saccharomyces cerevisiae.* Although conceptually simple, the algorithm efficiently extracted the upstream regulatory sequences that had been previously been determined experimentally for most of the yeast regulatory families already analyzed. Other studies [4-8,10-12] on the prediction of gene regulatory sequences have been based on oligonucleotide analysis.

Hsieh *et al.* [1] investigated the oligonucleotide distributions of typical microbial genomes and found that the microbial genomes have the statistical characteristics of a much shorter DNA sequence. This peculiar property supports an evolutionary model in which a genome evolves by random mutation but grows primarily by random segmental duplication.

Repetitive elements, including LINEs, SINEs, LTR and Alu, can be investigated in evolution analysis [2]. It has been estimated that at least 43% of the human genome is occupied by four major classes of interspersed repetitive elements – LINEs, SINEs, LTR elements and DNA transposons [2]. Their analysis has yielded some insights into the evolution of the human genome. The tandem repeats provided in our database can be used for forensic analysis and the study of genetic diseases [13-16].

Another application of the established database is to facilitate the design of primers to amplify specific regions of the genomic sequence. The basic concept is that the sequences of primers from 15 to 40 bps should be unique, unlike the repetitive oligonucleotides in our database. Additionally, our database maintains the repetitive oligonucleotides that facilitate the design of oligonucleotide probes to allow the selection of signature oligonucleotides when identifying different organisms using DNA arrays [9]. The user can query oligonucleotides whose lengths exceed a threshold, such as 15 bps, to determine whether the oligonucleotides are repetitive. The non-repetitive regions of the target sequences without repetitive oligonucleotides can be used as the signatures for the target genomes.

Construction and content

Data sources and contents

The proposed database provides information about sequence features generated by exhaustively computing

Table 2: Number of occurrences of the repetitive oligonucleotides in yeast genome

Repetitive oligonucleotide	Amount of occurrences			Repetitive oligonucleotide	Amount of occurrences		
	Genome	Up-streams	Coding region		Genome	Up-streams	Coding region
ACCCTA	2,724	822	793	CAATCCA	1,895	655	343
ACCCTC	2,917	881	795	CGTCTCC	592	199	148
AGTACT	3,073	933	879	CGTCTGA	652	196	165
AGTAGA	6,673	1,970	1,798	ACAAACTA	594	179	183
AGTAGC	4,912	1,545	1,299	ACAAACTC	514	175	112
GATACC	4,829	1,638	1,005	CACAGAAAC	146	38	46
GATAGA	7,030	2,163	1,807	CACAGAAGA	164	57	39
TGGTAA	10,513	3,493	2,214	ACATATAAAAA	54	9	29
TGTAAA	11,364	3,439	3,418	ACATATAAAAC	139	34	56
AAGGGGA	1,172	299	421	ACATATAAAAG	36	7	22
AAGGGGC	626	142	256	ACTTATGTCATC	57	17	23
AGAGTGG	983	310	271	ACTTCTAGTATA	159	44	67
AGAGTTA	1,859	610	441	ACTTTTTTTTCT	32	5	21
CAATCAG	1,358	445	320	ACTTTTTTTTTC	50	6	33

the sequences of the complete genome. The data sources including the complete genomes and the gene annotation information are obtained from GenBank [17]. The repetitive elements in the eukaryotic genomes, such as LINEs, SINEs, Alu and LTR, are obtained from Repbase [18]. The database supports a range of complete genomes including human, yeast, worm, and 128 microbial genomes. The Appendix lists the organisms supported in the database [see additional file 1].

The occurrences and the frequencies of oligonucleotides from one to 50 base pairs are generated and accumulated from each of the complete genome sequence. Inputting the sequence of the oligonucleotide returns the positions of the oligonucleotides. Additionally, both the occurrences and the frequencies of the repetitive elements such as LINEs, SINEs, Alu and LTR are provided by computationally scanning whole genome sequences. The tandem repeats are computationally detected by the tandem repeat finder [19]. The database also provides the gene annotation information. For instance, Table 2 presents the number of occurrences of repetitive oligonucleotides in yeast. The oligonucleotide "ACCCTA" occurs 2,724 times in the yeast genome, 822 times upstream of a gene (-600 ~-1 bp, +1 bp denotes the gene translational start postion) and 793 times in the coding regions. The counts of the occurrence of ecah oligonucleotide between one and 50 bps are present in the database.

Data Generation

A software is implemented to index systematically a complete genome sequence into a suffix-array using a perfect match approach [20]. This index is only able to find the perfect match for any oligonucleotide. The user can thus use it to find the positions of a designated oligonucleotide in a genome sequence. For each genome, the occurrences of all oligonucleotides shorter than 50 bps can be efficiently searched for. The occurrence is the position of the oligonucleotide in genome. The frequency is the count of oligonucleotide occurrences in a region. The regions are the complete genome, the coding regions and the non-coding regions. Frequencies of oligonucleotides with different lengths are stored in different flat-files. For example, the two chromosomes of the *Vibrio cholerae* genome are processed separately to allow the computation of the occurrences of oligonucleotides in each chromosomal sequence. RepeatMasker [21] and the repetitive element database, Repbase [18], are used to search the instances of the repetitive elements in eukaryotic genomes. The tandem repeat finder (TRF) is used to find the tandem repeats in genomes [19]. The TRF and RepeatMasker can find the instances of repetitive elements with imperfect matches. The settings used here for each software is described in below.

The Tandem Repeat Finder uses seven parameters. These are match score, mismatch score, indel score, probability of match and insertion, minimum score of alignment and the maximum of tandem repeat pattern size. The corresponding values used here are 2, 7, 7, 80, 10, 20 and 500. The values are the default suggestions found in Tandem Repeat Finder documentation.

The transposable elements (TEs) are detected by RepeatMasker. TEs in each genome are identified using the complete dataset available from REPbase Updates [Please add the citation]. The sensitivity and the speed of RepeatMasker are set as the default values.

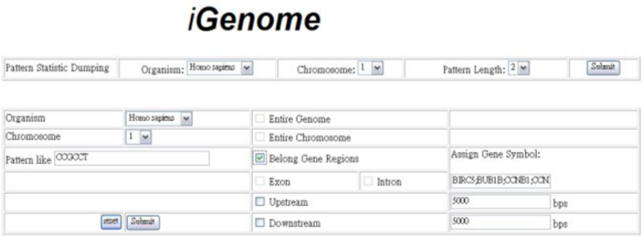
Table 3: Output styles of the database.

Database entries	Entry types	Output formats
Oligonucleotide occurrences	Positions	Web interface
Oligonucleotide frequencies	Counts	Web interface and flat-file
Gene coding regions	Positions	Web interface
Repetitive element frequencies (LINE, SINE, Alu, and so on)	Counts	Web interface and flat-file
Repetitive element occurrences	Positions	Web interface
Tandem repeats	Positions	Web interface

Utility

Table 3 presents the two output formats – flat-files and the web query interface with a filtering function. In the flat-file format, the fields of each oligonucleotide (in a chromosome) include the sequences, the number of occurrences in the chromosome, the number of occurrences in the coding regions and the number of occurrences in the non-coding regions. The user that requires a large amount of such data can download them in this format [1].

The web interface enables users to query the occurrences of an oligonucleotide in a genome and the number of occurrences in each chromosome. Figure 1 shows this approach. The occurrences of the repetitive elements and the tandem repeats in the established database can also be queried via the web interface, as in the example given in Fig 2. Figure 3 depicts the flat-file format of oligonucleotide frequencies. The first row in the flat-file presents the basic information for the oligonucleotide frequencies and the fields are the chromosome sequence/NCBI accession number, the length of the chromosome sequence, the size of coding regions, the size of non-coding regions, the length of the oligonucleotides and the minimum number of copies of the oligonucleotide. The directories labeled C10 are the files that contain the counts of oligonucleotides with at least ten occurrences in genome. Each file name includes the sequence/NCBI accession number, the length of oligonucleotides and the minimum occurrences of oligonucleotides. For example, "NC 000913 L30 C10" is the oligonucleotides, which are 30 nucleotides in length and have at least 10 occurrences in the genome.

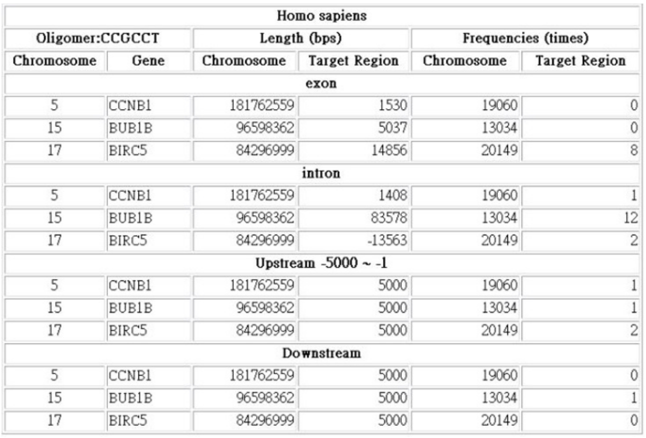


iGenome

Pattern Statistic Dumping	Organism: Homo sapiens	Chromosome: 1	Pattern Length: 2	Submit

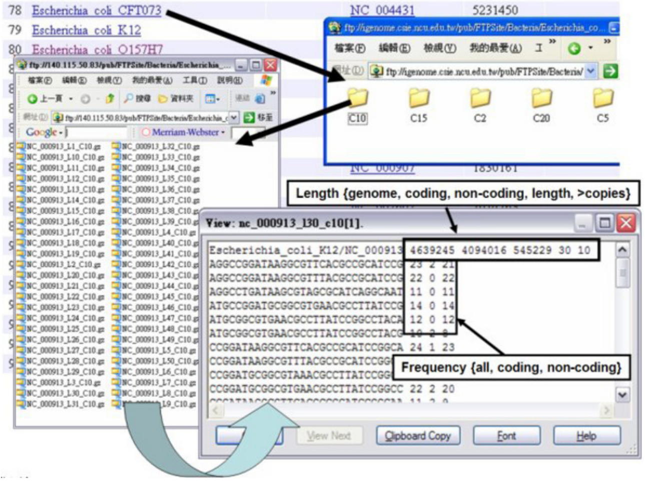
Organism	Homo sapiens	□ Entire Genome	
Chromosome	1	□ Entire Chromosome	
Pattern like CCGCCT		☑ Belong Gene Regions	Assign Gene Symbol:
		□ Exon □ Intron	BIRC5;BUB1B;CCNB1;CCN
		□ Upstream	5000 bps
reset	Submit	□ Downstream	5000 bps

Figure 1
Web query interface (1/2).



Homo sapiens					
Oligomer:CCGCCT		Length (bps)		Frequencies (times)	
Chromosome	Gene	Chromosome	Target Region	Chromosome	Target Region
exon					
5	CCNB1	181762559	1530	19060	0
15	BUB1B	96598362	5037	13034	0
17	BIRC5	84296999	14856	20149	8
intron					
5	CCNB1	181762559	1408	19060	1
15	BUB1B	96598362	83578	13034	12
17	BIRC5	84296999	-13563	20149	2
Upstream -5000 ~ -1					
5	CCNB1	181762559	5000	19060	1
15	BUB1B	96598362	5000	13034	1
17	BIRC5	84296999	5000	20149	2
Downstream					
5	CCNB1	181762559	5000	19060	0
15	BUB1B	96598362	5000	13034	1
17	BIRC5	84296999	5000	20149	0

Figure 2
Web query interface (2/2).



78 Escherichia coli CFT073 NC 004431 5231450
79 Escherichia coli K12
80 Escherichia coli O157H7

ftp://igenome.csie.ncu.edu.tw/pub/FTPSite/Bacteria/Escherichia_co...

ftp://140.115.50.83/pub/FTPSite/Bacteria/Escherichia_...

```
NC_000913_L1_C10.gz    NC_000913_L32_C10.gz
NC_000913_L10_C10.gz   NC_000913_L33_C10.gz
NC_000913_L11_C10.gz   NC_000913_L34_C10.gz
NC_000913_L12_C10.gz   NC_000913_L35_C10.gz
NC_000913_L13_C10.gz   NC_000913_L36_C10.gz
NC_000913_L14_C10.gz   NC_000913_L37_C10.gz
NC_000913_L15_C10.gz   NC_000913_L38_C10.gz
NC_000913_L16_C10.gz   NC_000913_L39_C10.gz
NC_000913_L17_C10.gz   NC_000913_L4_C10.gz
NC_000913_L18_C10.gz   NC_000913_L40_C10.gz
NC_000913_L19_C10.gz   NC_000913_L41_C10.gz
NC_000913_L2_C10.gz    NC_000913_L42_C10.gz
NC_000913_L20_C10.gz   NC_000913_L43_C10.gz
NC_000913_L21_C10.gz   NC_000913_L44_C10.gz
NC_000913_L22_C10.gz   NC_000913_L45_C10.gz
NC_000913_L23_C10.gz   NC_000913_L46_C10.gz
NC_000913_L24_C10.gz   NC_000913_L47_C10.gz
NC_000913_L25_C10.gz   NC_000913_L48_C10.gz
NC_000913_L26_C10.gz   NC_000913_L49_C10.gz
NC_000913_L27_C10.gz   NC_000913_L5_C10.gz
NC_000913_L28_C10.gz   NC_000913_L50_C10.gz
NC_000913_L29_C10.gz   NC_000913_L6_C10.gz
NC_000913_L3_C10.gz    NC_000913_L7_C10.gz
NC_000913_L30_C10.gz   NC_000913_L8_C10.gz
NC_000913_L31_C10.gz   NC_000913_L9_C10.gz
```

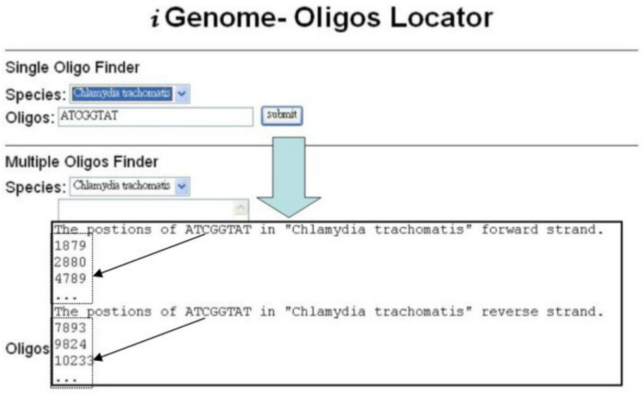
View: nc_000913_l30_c10[1].

```
Escherichia_coli_K12/NC_000913 4639245 4094016 545229 30 10
AGGCCGGATAAGGCGTTCACGCCGCATCCG 23 2 21
AGGCCGGATAAGGCGTTTACGCCGCATCCG 22 0 22
AGGCCTGATAAGCGTAGCGCATCAGGCAAT 11 0 11
ATGCCGGATGCGGCGTGAACGCCTTATCCG 14 0 14
ATGCGGCGTGAACGCCTTATCCGGCCTACA 12 0 12
ATGCGGCGTGAACGCCTTATCCGGCCTACG 10 2 8
CCGGATAAGGCGTTCACGCCGCATCCGGCA 24 1 23
CCGGATAAGGCGTTTACGCCGCATCCGG
CCGGATGCGGCGTAAACGCCTTATCCGG
CCGGATGCGGCGTGAACGCCTTATCCGGCC 22 2 20
```

Length {genome, coding, non-coding, length, >copies}

Frequency {all, coding, non-coding}

Figure 3
Database entries in flat-file format.



*i*Genome- Oligos Locator

Single Oligo Finder

Species: Chlamydia trachomatis

Oligos: ATCGGTAT

Multiple Oligos Finder

Species: Chlamydia trachomatis

Oligos

```
The postions of ATCGGTAT in "Chlamydia trachomatis" forward strand.
1879
2880
4789
...
The postions of ATCGGTAT in "Chlamydia trachomatis" reverse strand.
7893
9824
10233
...
```

Figure 4
The occurrence positions of the oligonucleotide are found by Oligos Locator.

Figure 4 shows the web interface for the occurrences of a specific oligonucleotide. The user submits the query oligonucleotide and selects particular species; the positions of the oligonucleotide are then shown. The first line is the user submitted data. Following this information are the positions of oligonucleotides in forward strand. This is followed by the same information for the reverse strand.

Conclusions

We have constructed the databases of both the oligonucleotide occurrence locations and their frequencies in the coding and the non-coding regions in complete genomes. The data in flat-file format can be downloaded directly for further analyses in several biological applications. The user may also use the web interface to query and access the database contents. The database also provides a filtering function for retrieving the information about oligonucleotides under search conditions specified by the users. Furthermore, the database provides the occurrences and the frequencies of other repetitive elements, such as LINE, SINE, Alu and tandem repeats in genomes.

Availability and requirements

The database is now available at http://igenome.csie.ncu.edu.tw/

Authors' contributions

FML implements the software and refinements the system. HDH conceives of the study and drafted the manuscripts. YCC and JTH participates the design and coordination. All authors read and approved the final manuscripts.

Additional material

Additional File 1
Appendix listing the organisms supported in the database.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-5-78-S1.doc]

Acknowledgement

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC93-2213-E-008-024.

References

1. Hsieh LC, Luo L, Ji F, Lee HC: **Minimal model for genome evolution and growth.** *Phys Rev Lett* 2003, **90:**18101.
2. Li WH, Gu Z, Wang H, Nekrutenko A: **Evolutionary analyses of the human genome.** *Nature* 2001, **409:**847-849.
3. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281:**827-842.
4. van Helden J, Andre B, Collado-Vides J: **A web site for the computational analysis of yeast regulatory sequences.** *Yeast* 2000, **16:**177-187.
5. Levy S, Hannenhalli S, Workman C: **Enrichment of regulatory signals in conserved non-coding genomic sequence.** *Bioinformatics* 2001, **17:**871-877.
6. Horng JT, Huang HD, Huang SL, Yan UC, Chang YC: **Mining putative regulatory elements in promoter regions of Saccharomyces cerevisiae.** In *Silico Biol* 2002, **2:**263-273.
7. Horng JT, Huang HD, Jin MH, Wu LC, Huang SL: **The repetitive sequence database and mining putative regulatory elements in gene promoter regions.** *J Comput Biol* 2002, **9:**621-640.
8. Huang HD, Chang HL, Tsou TS, Liu BJ, Kao CY, Horng JT: **A Data Mining Method to Predict Transcriptional Regulatory Sites Based on Differentially Expressed Genes in Human Genome.** *Third IEEE Symposium on BioInformatics and BioEngineering* Bethesda, Maryland, Computer Society, IEEE; 2003:297-304.
9. Kaderali L, Schliep A: **Selecting signature oligonucleotides to identify organisms using DNA arrays.** *Bioinformatics* 2002, **18:**1340-1349.
10. Horng JT, Lin FM, Lin JH, Huang HD, Liu BJ: **Database of repetitive elements in complete genomes and data mining using transcription factor binding sites.** *IEEE Trans Inf Technol Biomed* 2003, **7:**93-100.
11. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements in silico on a genomic scale.** *Genome Res* 1998, **8:**1202-1215.
12. van Helden J: **Regulatory sequence analysis tools.** *Nucleic Acids Res* 2003, **31:**3593-3596.
13. Lee JC, Cole M, Linacre A: **Identification of hallucinogenic fungi from the genera Psilocybe and Panaeolus by amplified fragment length polymorphism.** *Electrophoresis* 2000, **21:**1484-1487.
14. Lee JC, Cole M, Linacre A: **Identification of members of the genera Panaeolus and Psilocybe by a DNA test. A preliminary test for hallucinogenic fungi.** *Forensic Sci Int* 2000, **112:**123-133.
15. Koob MD, Moseley ML, Schut LJ, Benzow KA, Bird TD, Day JW, Ranum LP: **An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8).** *Nat Genet* 1999, **21:**379-384.
16. Cummings CJ, Zoghbi HY: **Fourteen and counting: unraveling trinucleotide repeat diseases.** *Hum Mol Genet* 2000, **9:**909-916.
17. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2003, **31:**23-27.
18. Jurka J: **Repbase update: a database and an electronic journal of repetitive elements.** *Trends Genet* 2000, **16:**418-420.
19. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.

20. Gusfield D: **Algorithm on Strings, Trees and Sequences.** NY, Cambridge University Press; 1997.
21. **RepeatMasker** [http://www.repeatmasker.org]

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."
Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral