

Sun Bio Workstation & Bio Cluster
Turning Innovation into value for the Life Science Industry

Technical White Paper

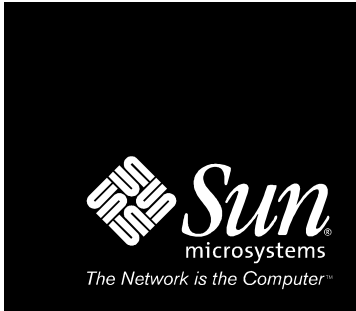


Table of Contents

Introduction.....	3
<i>Section A</i>	
Hardware & Software from Sun Microsystems.....	7
<i>Section B</i>	
Life Science Application Software.....	9
<i>Section C</i>	
Terms.....	11
<i>Section D</i>	
References.....	13
<i>Section E</i>	
Bio Cluster Reference Architecture.....	14
<i>Section F</i>	
Bio-ClusterGrid Applications List.....	15

Introduction:

slated as the next big opportunity for India to shine in the global arena after the success of IT services, Life Science research is gaining attention from industry , academic and government insitutions.

Several factors are contributing to the promotion of Life Science research in India

- Several multinationals have recognised India as a source of talent in the research and have set up facilities in India.
- As a WTO signatory, Indian pharma players face a do-or-die situation post December 2004 to either get relegated to a 'contract' or 'generics' manufacturers or strive to own IP on new drugs.
- Several Indian pharmaceutical companies have matured from playing 'generics' manufacturers to new drug makers with tremendous investments into research.
- Government backed Life Science research facilities have stepped up focus on modern, computer aided research in several areas – from genetics to system biology to CADD.

There are a growing demand for 'industry-ready' students by these research entities. Academic institutions are striving hard to fullfil this demand by promoting new courses . The critical need is to provide the research entities with sufficient and trained manpower to be productive on employment.

Sun in Computational Biology and Life Sciences

<http://www.sun.com/products-n-solutions/edu/commofinterest/compbio/>

<http://www.sun.com/lifesciences/>

Sun Microsystems has been a key technology provider to the life sciences industry for more than a decade.

This white paper & promotion highlights new technologies that can and are already benefiting the life science community. For example, Sun-AMD Opteron servers have recently set benchmark records for price *and* performance running life science (and other) applications. This white paper also focuses on existing technologies that offer cost saving and enhance productivity.

Sun's strength is not only in high performance computing with new leadership in high throughput computing, but in our numerous deep and

India aims to become the main bioinformatics hub ".....signs of a new star (ie. Biotechnology) rising on India's industrial horizon are already there"

***-Nature Biotechnology,
August 2004***

The recent unveiling of initiatives by Sun indicates India's growing strength in this area. Sun Microsystems, a global IT giant, has announced a Centre of Excellence (COE) for medical bioinformatics at Hyderabad, in collaboration with the Centre for DNA Fingerprinting and Diagnostics (CDFD) and the Government of Andhra Pradesh.

***-The Financial Express ,
January 2005 (India Brand
Equity Foundation)***

broad partnerships that we have developed along the way. Sun has more than 14,000 software vendor partners where some of the larger ones include AMD, Fujitsu, Oracle, SAS, Documentum, OpenText and Microsoft, which support multiple industries but have a special focus on life science.

In addition, we have more than 400 software partners focused solely on life science, along with 22 academic Centers of Excellence (COE) engaged in life science research (<http://www.sun.com/products-n-solutions/edu/programs/coe/>) projects ranging from systems biology to virtual surgery and bioinformatics on a super-computing scale.

Regular meetings with our COE and Computational Biology Special Interest Group : CB - SIG (<http://www.sun.com/products-n-solutions/edu/commofinterest/compbio/sig/>) members are held around the globe to foster collaboration within our community (see References in Section D to learn how to join).

The life science community is very comfortable with open source codes, while Sun is equally comfortable with contributing to and supporting these efforts. For example, Sun has recently donated Java3D to the open source community which is expected to have profound impact on the life science community given the rich visualization needs of this industry. And, few are aware that we are the second largest contributor of Linux code, behind only University of California Berkeley, and continue to donate millions of lines of code, as well as financial assistance to efforts that increase commonality between Linux and Solaris.

Sun is focused on ensuring that both open source and commercial life science applications in greatest demand are optimized for deployment on our platform. Sun offers excellent engineering support and dedicated annual programs, including Tunathon, which was established five years ago with the purpose of exposing and removing performance bottlenecks at the level of the source code. Modified code is then posted on the public domain (for example, BLAST, HMMER, etc.) or incorporated into the next release of the application for much improved performance.

Sun and Asia Pacific Science and Technology Center

<http://apstc.sun.com.sg/>

Popularly called as the APSTC, this is a joint collaboration between Sun Microsystems and the Nanyang Technological University, Singapore. APSTC aims to understand the computational requirements of various educational and research institutions. Simultaneously, APSTC's researchers and engineers are undertaking developed core research, applied research and application development in various high performance computing areas (grid computing, parallel computing, cluster computing, distributed computing, application optimisation & performance tuning). These skills are applied to various vertical segments of research like computational biology, electronic design and automation (EDA), financial analysis to help educational and research institutions in Asia-Pacific to help solve their computational needs.

"According to Carl Dahlman, manager of the Knowledge for Development Program at the World Bank Institute: "To benefit from the knowledge revolution, these three large countries will need to devise concrete strategies to address the four pillars of the knowledge economy: economic incentive and institutional regime, education and training, information infrastructure, and innovation system."

Read "Brazil, China and India Share Knowledge Strategies"., Wilton

Park (UK), 17 April 2001.

http://www.oecd.org/document/9/0,2340,en_2649_34269_2373065_1_1_1_1,00.html

"Sun™ ONE Grid Engine, Enterprise Edition enabled us to meet our peak demand turnaround requirements and run algorithms that simply weren't possible before."

– Alan Hart, Oxford Glycosciences, PLC

Solution:

Sun Microsystems and Asia Pacific Science & Technology Center – APSTC have put together a comprehensive solution combining Sun’s x86 based servers & workstations and 23 leading open-source based software and 3 popular development tools provided by APBionet. Called the Bio-Workstation & Bio-Cluster, the hardware and OS is detailed in Section A.

“Combining massive computing power and more traditional laboratorybased technology, this collaborative effort corresponds to a major breakthrough that is likely to revolutionize the discovery of antibiotics in this decade”

**–Mohammad Afshar,
RiboTargets Ltd.**

Bio-Workstation

- Sun Java Workstation w2100 with 64-bit Opteron processors & Linux
- Life Science Application software from APBionet (detailed in Section B)

Bio-Cluster

- Sun v20z servers with 64-bit Opteron processors with Linux as computational nodes
- Sun v40z server with 64-bit Opteron processors, Linux OS as Portal & NFS server.
- Sun Java Workstation w1100 with 64-bit Opteron processors, Linux and 3-D graphics from nVidia.
- Life Science Application software compiled by APBionet (detailed in Section B & F)
- Sun N1 Grid Engine Software
- Sun Java Enterprise System Sun’s Portal Server software.

Please refer to Bio Cluster Reference Architecture in Section E

Solution Benefits to Academia & Research Institutions :

The idea behind the Bio-Cluster & Bio-Workstation is to allow biologists to run bioinformatics applications on high performance cluster and workstation in the easiest possible manner so that they can focus their precious time and efforts to undertake biological experiments.

Overall, I really like the W2100z. Graphics performance was great, while maintaining the fast band width of a server. This machine would be great for applications that require moving a lot of data, graphic or otherwise.

**–George Maestri , Computer
Graphics World**

- **One-stop integrated solution** featuring the popular open-source Life Science software and Sun’s high-performance servers and workstations.
- **Saves time & complexity** by avoiding hours of download and configuring the software from various sources.
- **Ease-of-use** through the portal interface that provides a web based window to all the applications.

- *23 leading pre-compiled applications and 3 popular development tools* that have been tested & optimized for the computing environment suggested.
- *Resource optimisation & virtualisation* using grid technologies as provided by Sun N1 Grid Engine 6.0.
- *Open standards based open source software* are provided including the N1 Grid Engine.
- *Reference architecture based* on testing & evaluation at the APSTC.

Commercials :

The commercials of the solution consists of

- Sun servers and workstations
- Software licenses and media
- Sun services for installation, training & integration of the solution.
- Sun support charges for Sun hardware & software
- Third party products that need to be sourced from Sun Authorized partners.

The end customer pricing on the promotion are available from Sun's Authorised Partners.

See Section C for Terms of offer.

"We will be using the Sun Java Workstation W2100z on AMD Opteron processors for research on robotics, security, digital mapping and any projects requiring 64-bit computing," "We believe the Sun Java Workstation W2100z is extremely versatile. Its single architecture will enable us to support simultaneous 32-bit and 64-bit computing with no compromises in performance. It will allow us to maintain our existing x86 infrastructure while enabling a smooth migration to next-generation 64-bit operating systems and applications when required. We look forward to continued innovation from Sun and AMD."

*-David Livingston
associate director of computer science research, Carnegie Mellon University.*

Section A

Hardware & Software from Sun Microsystems

Bio-Workstation

- 1 x *Sun Java Workstation w2100* with
 - 2 x AMD Opteron 250 processors
 - 4 GB RAM
 - nVidia FX 3000 graphics
 - Sun 17" Monitor
- Red Hat 3.0 Workstation OS (licensed)
- Life Science Application software as in Section B & F

Bio-Cluster

Head Node

- 1 x *SunFire v40z* server with
 - 2 x AMD Opteron 250 processors
 - 4 GB RAM
- Red Hat 2.1 server Linux OS (licensed)
- Sun Java System Portal Server for Linux (licensed)
- Sun N1 Grid Engine Software (master)

Execution Nodes

- 3 x *SunFire v20z* server with
 - 1 x AMD Opteron 250 processors
 - 2 GB RAM
- Fedora Core 2 server OS (licensed)
- Sun N1 Grid Engine Software (client)
- Life Science Application software as in Section B & F

Visualisation & Access

- 1 x *Sun Java Workstation w1100* with
 - 1 x AMD Opteron 250 processors
 - 2 GB RAM
 - nVidia FX 500 graphics
- Red Hat 3.0 Workstation OS (licensed)
- 17" Sun Monitor

Section B

Life Science Application Software

Application Area	Application
Homology & Similarity Search	Blast- Fasta GlimmerM Wise2
Sequence Analysis	ACT ClustalW EMBOSS HMMER Image T-Coffee
Structural Prediction	FastDNAmI LOOPP MapMaker QTL PAML PHYLIP

Application Area	Application
Molecular Modeling	Artermis Cn3D GROMACS RasMol ReadSeq TribeMCL NAMD NMRView
Development Tools	BioJava BioPerl BioPython

For a detailed note on all the applications & tools, please go to Section F.

Disclaimer :

'All open source applications & modelling software are packaged and distributed by APBioNet. Sun Microsystems merely provides the scripts to integrate these applications with NIGE and the operating systems. These open source applications are owned by their respective owners'.

Section C

Terms

Bio-Workstation

Hardware and Software

- 1 x *Sun Java Workstation w2100* with
 - 2 x AMD Opteron 250 processors
 - 4 GB RAM
 - nVidia FX 3000 graphics
- Red Hat 3.0 workstation OS (licensed)
- Sun 17" Monitor
- Application software as in Section B

Warranty

- | | |
|---------------------------------|-----------------------------|
| · Sun Java Workstation w2100 | 1 year standard warranty |
| · Red Hat 3.0 Workstation OS | 90 days media warranty only |
| · Application s/w thru APBionet | None |

Services from APSTC / Sun Partner

- Configuration & installation of the Bio Workstation
- Configuration & installation of the Life Science Applications bundled

Bio-Cluster

Head Node

- 1 x *SunFire v40z server* with
 - 2 x AMD Opteron 250 processors
 - 4 GB RAM
- Red Hat 2.1 server OS (licensed)

- Sun Java System Portal Server for Linux – 2 CPU License
- Sun N1 Grid Engine Software – 50 CPU License

Execution Node

- 3 x *SunFire v20z server* with
 - 1 x AMD Opteron 250 processors
 - 2 GB RAM
- RedHat 3.0 AS (licensed)
- Sun N1 Grid Engine Software (client)
- Application software as in Section B

Visualisation & Access

- 1 x *Sun Java Workstation w1100* with
 - 1 x AMD Opteron 250 processors
 - 2 GB RAM
 - nVidia FX 500 graphics
- Red Hat 3.0 Workstation OS (licensed)
- 17" Sun Monitor

Warranty

- | | |
|---------------------------------|-----------------------------|
| • Sun Fire servers v20 , v40 | 3 year standard warranty |
| • Sun Java Workstation w1100 | 1 year standard warranty |
| • RedHat OS | 90 days media warranty only |
| • Sun Java System Portal Server | 90 days media warranty only |
| • Application s/w thru APBionet | None |

Services from APSTC / Sun Authorised Partners (2 Man Days)

- Configuration & installation of the Bio Workstation
- End user training to use the Bio-ClusterGrid portal
- Administrator training (manage users, add applications to the portal etc.)
- Basic Grid Training (basic usage of Sun Grid Engine)

Note

- Prices are available from Sun Authorized partners.
- Half-height rack , cabling & 12 port Gigabit Ethernet Switch to be supplied by Sun partner.
- Optional Spectrum Sun Support upgrades are available on request

Section D

References

1. Sun in Education & Research
<http://www.sun.com/edu>
2. Sun in HPC
<http://www.sun.com/hpc>
3. Sun in Life Sciences
<http://www.sun.com/lifesciences/>
4. Sun Computational Biology - Special Interest Group
http://www.sun.com/products-n-solutions/edu/commofinterest/_compbio/sig/
5. Sun in Computational Biology
<http://www.sun.com/products-n-solutions/edu/commofinterest/compbio/>
6. Sun Java Workstations
<http://www.sun.com/desktop/products/ws.html>
7. Sun Fire Entry Servers
<http://www.sun.com/servers/entry/>
8. Sun Grid Engine
<http://www.sun.com/software/gridware/>
9. Sun Java Enterprise Portal Server
http://www.sun.com/software/products/portal_srvr/home_portal6.html
10. Asia Pacific Science & Technology Center
<http://apstc.sun.com.sg>
11. Sun EduConnection Newsletter
<http://www.sun.com/products-n-solutions/edu/educonnection/>

Downloads

1. Sun's Computational Biology Whitepaper
http://www.sun.com/products-n-solutions/edu/whitepapers/pdf/SunBriefing_compbio.pdf

Section E

BIO CLUSTER REFERENCE ARCHITECTURE



Section F

Bio-ClusterGrid Applications List

APPLICATION	DESCRIPTION	USAGE	CONTACT
Homology & Similarity Search			
Biodas(dazzle)	Distributed Annotation System(DAS) is a server system for the sharing of Reference Sequences. DAS relies on there being a common "reference sequence" on which to base annotations. The references sequence consists of a set of "entry points" into the sequence. http://biodas.org	<u>Deployment Guide</u> http://www.biojava.org/dazzle/deploy.html	Author Scott Pearson Mailing List das@biodas.org
BLAST	BLAST (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. BLAST uses a heuristic algorithm that seeks local as opposed to global alignments and is therefore able to detect relationships among sequences that share only isolated regions of similarity (Altschul et al. 1990). For a better understanding of BLAST, you can refer to the BLAST Course, which explains the basics of the BLAST algorithm. There is also a description of BLAST services located here. Also for details on BLAST and theory of similarity search, see the References section .http://www.ncbi.nlm.nih.gov/BLAST		Author Stephen Altschul, Jonathan Epstein, David Lipman, Tom Madden, Scott McGinnis, Jim Ostell, Alex Schaffer, Sergei Shavirin, Heidi Sofia, Jinghui Zhang Email blast-lp@ncbi.nlm.nih.gov

APPLICATION	DESCRIPTION	USAGE	CONTACT
Homology & Similarity Search			
Fasta	<p>FASTA@ (pronounced FAST-Aye) stands for FAST-All, reflecting the fact that it can be used for a fast protein comparison or a fast nucleotide comparison. This program achieves a high level of sensitivity for similarity searching at high speed. This is achieved by performing optimised searches for local alignments using a substitution matrix. The high speed of this program is achieved by using the observed pattern of word hits to identify potential matches before attempting the more time consuming optimised search. The trade-off between speed and sensitivity is controlled by the ktup parameter, which specifies the size of the word. Increasing the ktup decreases the number of background hits. Not every word hit is investigated but instead initially looks for segment's containing several nearby hits.</p> <p>ftp://ftp.virginia.edu/pub/fasta/</p>	<p>fasta34 musplfm.aa prop_test.lib</p>	<p>Author William R. Pearson Mailwrp@virginia.edu</p>
GlimmerM	<p>A gene finder derived from Glimmer, but developed specifically for eukaryotes. It is based on a dynamic programming algorithm that considers all combinations of possible exons for inclusion in a gene model and chooses the best of these combinations. The decision about what gene model is best is a combination of the strength of the splice sites and the score of the exons generated by an interpolated Markov model (IMM). The system has been trained for Arabidopsis thaliana, Oryza sativa (rice), and Plasmodium falciparum (the malaria parasite), and should work well on closely related organisms. Use the GlimmerM Web Interface to run GlimmerM directly, or see below for instructions on downloading the complete system including source code.</p> <p>http://www.tigr.org/software/glimmer/</p>	<p>glimmerm_sun <genome_file></p>	<p>Author http://www.tigr.org/tigr-scripts/mailler_scripts/contact_us.pl?receiver=mpertea&refer_text=GlimmerM</p> <p>Email</p>

APPLICATION	DESCRIPTION	USAGE	CONTACT
Sequence Analysis			
Wise	<p>Wise2 is a package focused on comparisons of bio polymers, commonly DNA sequence and protein sequence. There are many other packages which do this, probably the best known being BLAST package (from NCBI) and the Fasta package (from Bill Pearson). There are other packages, such as the HMMER package (Sean Eddy) or SAM package (UC Santa Cruz) focused on hidden Markov models (HMMs) of bio polymers. Wise2's particular forte is the comparison of DNA sequence at the level of its protein translation. This comparison allows the simultaneous prediction of say gene structure with homology based alignment. There is currently no other package that I know of that contains this type of algorithm with a full blown gene prediction model and a hidden Markov model of a protein domain. Wise2 also contains other algorithms, such as the venerable Smith-Waterman algorithm, or more modern ones such as Stephen Altschul's generalized gap penalties, or even experimental ones developed in house, such as dba (see section 7.1). The development of these algorithms is due to the ease of developing such algorithms in the environment used by Wise2. Wise2 has also been written with an eye for reuse and maintainability. Although it is a pure C package you can access its functionality directly in Perl. Parts of the package (or the entire package) can be used by other C or C++ programs without name space clashes as all externally linked variables have the unique identifier Wise2 prep ended. Java and CORBA ports are being considered – see 8 the API section</p> <p>http://www.ebi.ac.uk/Wise2/doc_wise2.html</p>	<p>genewise protein.pep cosmid.dnagenewisedb – proddb protein.pep –dnas cosmid.dna</p>	<p>Author Ewan Birney, Richard Copley</p> <p>Email birney@ebi.ac.uk, richard.copley@embl-heidelberg.de</p>
ACT	<p>ACT(Arternis Comparison Tool) is a DNA sequence comparison viewer based on Arternis. It can read complete EMBL and GENBANK entries or sequence in FASTA or raw format. Extra sequence features can be in EMBL, GENBANK or GFF format</p> <p>http://www.sanger.ac.uk/Software/ACT/v2/</p>	<p><code>./act etc/af063097.embl etc/ af063067_v_b132222.crunch etc/af063097.fasta</code></p>	<p>Author K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M-A. Rajandream and B. Barrell</p> <p>Email artemis@sanger.ac.uk</p>
ClustalW	<p>ClustalW is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.</p> <p>http://www.ebi.ac.uk/clustalw/</p>	<p><code>./act etc/af063097.embl etc/ af063067_v_b132222.crunch etc/af063097.fasta</code></p>	<p>Author Julie Thompson and Francois Jeanmougin</p> <p>Mail julie@igbmc.u-strasbg.fr</p>

APPLICATION	DESCRIPTION	USAGE	CONTACT
Sequence Analysis			
ClustalW	<p>Emboss is a software analysis package specially developed for the needs of the molecular biology user community. It automatically copes with data in a variety of formats and even allows transparent retrieval of sequences data from the web.</p> <p>http://www.hgmp.mrc.ac.uk/Software/EMBOSS/download.htm</p>	<p>Full list of applications can be found at</p> <p>http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/index.html</p>	<p>Author http://www.hgmp.mrc.ac.uk/Software/EMBOSS/credits.html</p> <p>Mail emboss@embnet.org</p>
HMMER	<p>HMMER is an implementation of profile HMM methods for sensitive database searches using multiple sequence alignments as queries. Basically, you give HMMER a multiple sequence alignment as input; it builds a statistical model called a "hidden Markov model" which you can then use as a query into a sequence database to find (and/or align) additional homologues of the sequence family.</p> <p>http://hmmmer.wustl.edu/</p>	<p>1. Set path to hmmer binaries 2. cd to directory 'test'</p> <p><code>hmmsearch globin.hmm Artemia.fa</code></p>	<p>Author National Human Genome Research Institute</p> <p>Mail eddy@genetics.wustl.edu</p>
Image	<p>Image is a package of analysis algorithms for processing gel images from restriction digest fingerprinting experiments. Image has been tightly integrated with a friendly user interface and provides a robust tool for large scale physical mapping. Image is able to process gels from a wide variety of scanning technologies and has been tested on various fingerprinting protocols, producing normalized band and gel images as output. http://www.sanger.ac.uk/Software/Image/download.shtml</p>	<p>Run <code>im3</code> after importing the data</p>	<p>Author Fred Wobus, Richard Durbin, Simon Kelley, John Bradley</p> <p>Mail image@sanger.ac.uk</p>
T-Coffee	<p>T-Coffee is a multiple sequence alignment package. Given a set of sequences (Proteins or DNA), T-Coffee generates a multiple sequence alignment. T-Coffee allows the combination of a collection of multiple/ pairwise, global or local alignments into a single model. It also allows to estimate the level of consistency of each position within the new alignment with the rest of the alignments. See the pre-print for more information. T-Coffee Replaces The COFFEE objective function that had been implemented in SAGA. Please, note that COFFEE will not be maintained anymore in SAGA. We recommend to use T-Coffee instead http://igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/t_coffee_home_page.html</p>	<p><code>t_coffee -in fast_pair test.peport_coffee arp.pep</code></p>	<p>Author Alexandre Gattiker, Liisa Holm, Laurent Falquet, Anne-List Veuthey, Amos Bairoch</p> <p>Mail cedric.notredame@europa.com</p>

APPLICATION	DESCRIPTION	USAGE	CONTACT
Structural Prediction			
FastDNAMl	<p>astDNAMl is a program for estimating maximum likelihood phylogenetic trees from nucleotide sequences.</p> <p>http://geta.life.uiuc.edu/~gary/programs/fastDNAMl.html</p> <p>(MPI Version Available)</p>	<p>Set PATH to access the fastDNAMl binaryf</p> <p>astDNAMl < test5.phy > test5.out</p>	<p>Author Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. 1994.</p> <p>Mail gary@phylo.life.uiuc.edu</p>
LOOPP	<p>LOOPP is a program for PROTEIN RECOGNITION and design of PROTEIN FOLDING potentials. LOOPP (Learning, Observing and Outputting Protein Patterns) evolved from our previous LOOPP (Linear Optimization of Protein Potentials). As suggested by the new/old name there is some continuity, but there are also many new features. LOOPP performs sequence to sequence, sequence to structure (threading), and structure to structure alignments. It further enables the optimization of potentials and scoring functions for the above mentioned applications. One may also use LOOPP to generate non-redundant libraries of folds, both for the training and recognition'</p> <p>.http://cbsu.tc.cornell.edu/software/loopp/index.htm</p>	<p>To start using the program one needs basically three things: a) file with sequences (with the default name SEQ) b) file with coordinates of residues in reduced representation (XYZ) c) file with query sequences for recognition (with the default name seq_to_examine.txt) loop - q -p hl_eij.mat -m -te -o hl.log</p>	<p>Author J. Meller and R.Elber</p> <p>Mail meller@cs.cornell.edu</p>
MapMaker/QTL	<p>MAPMAKER is a linkage analysis package designed to help construct primary linkage maps of markers segregating in experimental crosses. MAPMAKER performs full multipoint linkage analysis (simultaneous estimation of all recombination fractions from the primary data) for dominant, recessive, and co-dominant (e.g. RFLP-like) markers. QTL is a companion program to MAPMAKER which allows one to map genes controlling polygenic quantitative traits in F2 intercrosses and BC1 backcrosses relative to a genetic linkage map.</p> <p>ftp://ftp-genome.wi.mit.edu/distribution/software/newqtl/</p>	<p>qtl.sunType help to get all the available commands quit to exit</p>	<p>Author Leonid Kruglyak ,Mark Daly</p> <p>Mail leonid@genome.wi.mit.edu edumjdaly@genome.wi.mit.edu</p>
PAML	<p>PAML is a program package for phylogenetic analyses of DNA or protein sequences using maximum likelihood. PAML may be useful if you are interested in the process of sequence evolution. The two main programs, baseml and codeml, implement a number of sophisticated models, which you can use to construct likelihood ratio tests of evolutionary hypotheses. Right now, the following options/models do not seem available in other packages.</p> <p>http://abacus.gene.ucl.ac.uk/software/paml.html</p>	<p>Execute the individual executable with the appropriate input files given eg. baseml baseml.ctf</p>	<p>Author Yang Ziheng</p> <p>Mail http://www.rannala.org/gst/</p>

APPLICATION	DESCRIPTION	USAGE	CONTACT
Structural Prediction			
PHYLIP	<p>PHYLIP(the PHYLogeny Inference Package) is a package of programs for inferring phylogenies(evolutionary trees).The programs are controlled through a menu, which asks the users which options they want to set, and allows them to start the computation. The data are read into the program from a text file, which the user can prepare using any word processor or text editor (but it is important that this text file not be in the special format of that word processor – it should instead be in “flat ASCII” or “Text Only” format). Some sequence analysis programs such as alignment programs can write data files in the PHYLIP format. Most of the programs look for the data in a file called “infile” – if they do not find this file they then ask the user to type in the file name of the data file.Output is written onto special files with names like “outfile” and “treefile”. Trees written onto “treefile” are in the Newick format, an informal standard agreed to in 1986 by authors of a number of major phylogeny packages.</p> <p>http://evolution.genetics.washington.edu/phylip.html</p>	Executing of individual executable files using the appropriate database	<p>Author http://evolution.genetics.washington.edu/phylip/credits.html</p> <p>Mail joe@gs.washington.edu</p>
NAMD	<p>NAMD, recipient of a 2002 Gordon Bell Award, is a parallel, object-oriented molecular dynamics code designed for high-performance simulation of large biomolecular systems. NAMD scales to hundreds of processors on high-end parallel platforms and tens of processors on commodity clusters using switched fast ethernet. NAMD is file-compatible with AMBER, CHARMM, and X-PLOR and is distributed free of charge with source code.</p> <p>http://www.ks.uiuc.edu/Research/namd/</p>	charmrun namd2 ++local +p<procs> <configfile>	<p>Mail namd@ks.uiuc.edu</p>
Artemis	<p>Artemis is a free genome viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of the sequence, and its six-frame translation. Artemis is written in Java, and is available for UNIX, GNU/Linux, BSD, Macintosh and MS Windows systems. It can read complete EMBL and GENBANK database entries or sequence in FASTA or raw format. Extra sequence features can be in EMBL, GENBANK or GFF format.</p> <p>http://www.sanger.ac.uk/Software/Artemis/v5/</p>	art etc/c1215.embl	<p>Author K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M-A. Rajandream and B. Barrell</p> <p>Mail artemis-users@sanger.ac.uk</p>
Cn3D	<p>Cn3D is a helper application for your web browser that allows you to view 3-dimensional structures from NCBI's Entrez retrieval service. Cn3D runs on Windows, Macintosh, and Unix. Cn3D simultaneously displays structure, sequence, and alignment, and now has powerful annotation and alignment editing features.</p> <p>http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml</p>	<p>Configure your browser so that it knows where Cn3D is and that it should use Cn3D when it receives a file of type <code>chemical/ncbi-asn1-binary</code></p> <p>http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dinstall.shtml#browser</p>	<p>Author Stephen H. Bryant</p> <p>Mail bryant@ncbi.nlm.nih.gov</p>

APPLICATION	DESCRIPTION	USAGE	CONTACT
Structural Prediction			
GROMACS	<p>GROMACS is a versatile package to perform molecular dynamics, i.e. simulate the Newtonian equations of motion for systems with hundreds to millions of particles. It is primarily designed for biochemical molecules like proteins and lipids that have a lot of complicated bonded interactions, but since GROMACS is extremely fast at calculating the nonbonded interactions (that usually dominate simulations) many groups are also using it for research on non-biological systems, e.g. polymers</p> <p>.http://www.gromacs.org/download/index.php</p>	<code>cd gmxdemo./demo</code>	<p>Authors Berk Hess, Erik Lindahl, David van der Spoel</p> <p>Mail gmx-users@gromacs.org</p>
RasMol	<p>RasMol is a molecular graphics program intended for the visualisation of proteins, nucleic acids and small molecules. The program is aimed at display, teaching and generation of publication quality images. The program has been developed at the University of Edinburgh's Biocomputing Research Unit and the Biomolecular Structures Group at Glaxo Research and Development, Greenford, UK. RasMol reads in molecular co-ordinate files in a number of formats and interactively displays the molecule on the screen in a variety of colour schemes and representations. Currently supported input file formats include Brookhaven Protein Databank (PDB), Tripos' Alchemy and Sybyl Mol2 formats, Molecular Design Limited's (MDL) Mol file format, Minnesota Supercomputer Center's (MSC) XMol XYZ format, CHARMM format, MOPAC format, CIF format and mmCIF format files. If connectivity information and/or secondary structure information is not contained in the file this is calculated automatically. The loaded molecule may be shown as wireframe, cylinder (driending) stick bonds, alpha-carbon trace, spacefilling (CPK) spheres, macromolecular ribbons (either smooth shaded solid ribbons or parallel strands), hydrogen bonding and dot surface.</p> <p>http://www.bernstein-plus-sons.com/software/rasmol</p>	Download a sample PDB <code>filerasmol sample.pdb</code>	<p>Author Herbert J. Bernstein</p> <p>Mail yaya@bernstein-plus-sons.com</p>
ReadSeq	<p>Reads and writes nucleic/protein sequences in various formats. Data files may have multiple sequences. Readseq is particularly useful as it automatically detects many sequence formats, and interconverts among them. Formats added to this release include + MSF multi sequence format used by GCG software + PAUP's multiple sequence (NEXUS) format + PIR/CODATA format used by PIR + ASN.1 format used by NCBI + Pretty print with various options for nice looking output.</p> <p>http://iubio.bio.indiana.edu/soft/molbio/readseq</p>	<code>./readseq (C version)java -jar readseq.jar (Java version)</code>	<p>Author Don Gilbert</p> <p>Mail software@bio.indiana.edu</p>

APPLICATION	DESCRIPTION	USAGE	CONTACT
Structural Prediction			
TribemMCL	<p>TribemMCL is a method for clustering proteins into related groups, which are termed 'protein families'. This clustering is achieved by analysing similarity patterns between proteins in a given dataset, and using these patterns to assign proteins into related groups. In many cases, proteins in the same protein family will have similar functional properties. TribemMCL uses a novel clustering method (Markov Clustering or MCL) which solves problems which normally hinder protein sequence clustering.</p> <p>http://www.ebi.ac.uk/research/cgg/tribe/</p>	tribemcl	<p>Author Anton Enright</p> <p>Mail tribemcl@ebi.ac.uk</p>
Tools			
Biojava	<p>BioJava is a general bioinformatics toolkit. It provides a framework for building everything from simple scripts to complete applications. BioJava is designed to be used as a library, so to make it usable we must:</p> <ul style="list-style-type: none"> * Design by Interface but provide working implementations so that you can always extend or replace behavior and implementations. * Provide extensive API documentation as well as a clear overview of how it all fits together. * Give simple examples that show how to use the APIs. <p>http://www.biojava.org</p>	<p>Download the source file, unzip and untar. Do the following steps</p> <pre>cd demosjavac seq/ TestEmbl.java java seq.TestEmbl seq/ AL121903.embl</pre>	<p>Author Thomas Down</p> <p>Mail td2@sanger.ac.uk</p>
Bioperl	<p>Bioperl is a collection of perl modules that facilitate the development of perl scripts for bioinformatics applications. In order to take advantage of bioperl, the user needs a basic understanding of the perl programming language including an understanding of how to use perl references, modules, objects and methods.</p> <p>http://www.bioperl.org</p>	<p>The Bioperl test system is located in the t/ directory and is automatically run whenever you execute the 'make test' command. Alternatively if you want to investigate the behavior of a specific test such as the SeqIO test you would type:</p> <pre>% perl -I. -w t/SeqIO.t</pre>	<p>Author http://bioperl.org/Participants/</p> <p>Mail bioperl-l@bioperl.org</p>
Biopython	<p>Biopython is a set of libraries to provide the ability to deal with "things" of interest to biologists working on the computer. In general this means that you will need to have at least some programming experience (in python, or at least an interest in learning to program. Biopython's job is to make your job easier as a programmer by supplying reusable libraries so that you can focus on answering your specific question of interest, instead of focusing on the internals of parsing a particular file format.</p> <p>http://www.biopython.org</p>	<pre>pythonfrom Bio.Seq import Seqfrom Bio.Alphabet.IUPAC import unambiguous_dna as new_seq = Seq('GATCAGAAG', unambiguous_dna new_seq[0:2])from Bio import Translatetranslator = Translate.unambiguous_ dna_by_name["S tandard"]translator.tr anslate(new_seq)</pre>	<p>Author http://www.biopython.org/participants/</p> <p>Mail biopython@biopython.org</p>

Sun Microsystems India Private Limited, 6th Floor, Prestige Obelisk, No. 3, Kasturba Road, Bangalore 560 001, India
Tel: 91-80-5693 0600, Fax: 91-80-5693 0655/0666, Toll Free Line : 1-600-338072



©2005 Sun Microsystems, Inc. All rights reserved. Sun, Sun Microsystems, the Sun logo, Solaris, Sun Fire are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and other countries. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARCInternational, Inc. in the United States and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc. Information subject to change without notice.