SEQUEST® Cluster

An integrated solution – customized to accelerate protein identification

Analyze • Detect • Measure • Control™

Thermo ELECTRON CORPORATION

- Supercomputing capability with the ease-of-use of a single PC
- Parallel processing to reduce SEQUEST search times and accelerate protein analysis
- Remote operation through the cluster daemon
- Scalable configurations to meet every proteomics lab and budget
- Robust cluster systems management from a single point of control
- Linux™ and Windows® options available
- Three supported hardware vendors with customized configurations
- Includes all the new innovative tools that are part of BioWorks™ 3.1 software

Harnessing the Power of a Supercomputer

The SEQUEST Cluster offers an integrated, pre-configured software solution that enables accelerated processing of complex proteomics data generated by LC/MS^n experiments. Providing a perfect complement to systems such as the Finnigan ProteomeX,™ Finnigan LTQ,™ or Finnigan LTQ-FT,™ the Cluster allows users to immediately reap the benefits of parallel processing from a single desktop.

The Demand for Parallel Processing

Tandem mass spectrometry has become increasingly fast and more accurate, capable of generating thousands of spectra for complex mixtures in a single hour. In order to identify the proteins and their post-translational modifications present in each sample, the majority of these spectra must be processed through BioWorks software, with resulting database search times often exceeding an entire day. These search times can be compounded by many different parameters, such as the size of the protein database being searched, the number of spectra that were acquired, and the number of post-translational modifications that are being mined. An average ProteomeX dataset may take as long as three days to analyze on a single CPU! This volume of data necessitates true Mass Informatics capability, and the ability to search large datasets on more than just a single processor.

High-Throughput Database Searching in a Fraction of the Time

SEQUEST Cluster Configuration	No. of Processors	Analysis Time
Single PC	1	10 hours, 43 minutes
4-node Cluster	9	57 minutes
8-node Cluster	17	30 minutes
16-node Cluster	33	16 minutes

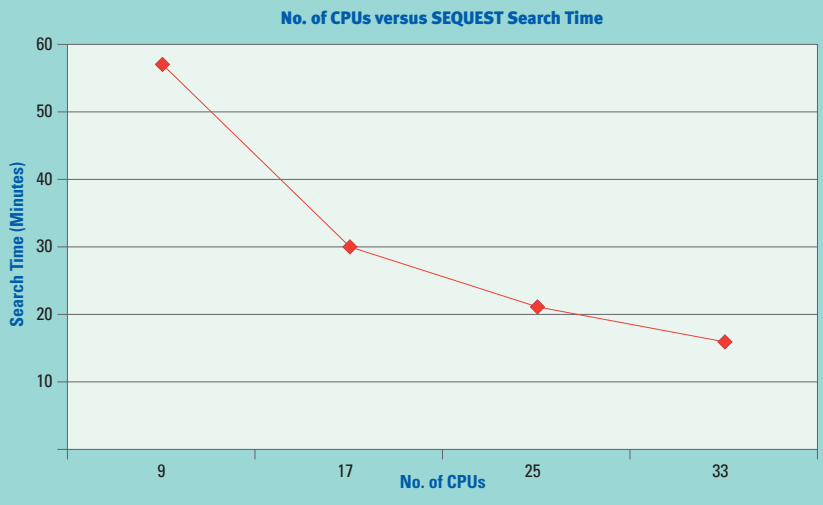
From over 10 hours of analysis time during database searching to only 16 minutes!

Benchmark Results for the SEQUEST Cluster

LTQ Datafile:	140-minute acquisition
Number of spectra:	17,958
Protein Database:	human_ref.fasta
Number of protein entries:	37,490

An example of the increased power that can be derived from the SEQUEST Cluster running on 4, 8 and 16 parallel processing nodes is illustrated in the above benchmark run using a data file of a mixture of human proteins acquired for over two hours, resulting in almost 18,000 spectra. The small protein database against which the file was searched included just over 37,000 entries. The benchmark search times show that a near 11-hour database search time can easily be reduced to approximately 16 minutes by adding an increasing number of dual-processor computing nodes, in addition to each dual-processor management node.

The SEQUEST Cluster dramatically increases the speed and capacity to identify unknown proteins by dividing the database search process into separate tasks and performing them in parallel. Once the computing nodes have completed each search, the results are integrated on the management node, and a single result output is generated.



This graph shows the scalability of the number of processors vs. SEQUEST search time for the above example.

Effortless Node Management and Remote Operation

Using the framework of **Parallel Virtual Machine (PVM)** programming, the SEQUEST Cluster can allow multiple processors on computing nodes to communicate with each other simultaneously.

The management node for the SEQUEST Cluster runs on the Microsoft® Windows 2000 Server with Service Pack 2 operating system, and contains the raw MS/MS data files used for the database searches, the fasta databases, and the .dta and .out files generated by SEQUEST.

The SEQUEST Cluster supports two operating systems for the computing nodes, each being a dual processor:

- **Microsoft Windows 2000 Professional, or**
- **Red Hat™ Linux, version 8.0**

Interaction with the management and computing nodes of the SEQUEST Cluster is performed through the BioWorks 3.1 protein identification software interface, as seamlessly as if the processing were being performed on a standard desktop PC.

The **SEQUEST Queue** enables unattended batch processing of database searches, enhancing productivity without the need for monitoring.

The new TurboSEQUEST™ Batch Search utility allows an entire sequence or list of samples from an acquisition to be processed.

The **Cluster Daemon** provides the ability to operate the SEQUEST Cluster remotely, by allowing users to drop files into a specified directory, along with a search parameter file of their choice, ***from any PC in the lab connected to the Cluster server.*** The Daemon regularly monitors the directory, and initiates the Cluster when raw files and search parameter files are present.

Scalable Performance to Match the Size of Any Budget

From as few as 4 dual-CPU nodes, to as many as 256 dual-CPU nodes, customizable sizing options allow users to start with a SEQUEST Cluster as small as their budget requires, and have the flexibility to scale up to meet the demands of their expanding labs. More computing nodes can be added at any time to handle increasing data workloads, which is why clustering offers significant price/performance advantages over purchasing additional SEQUEST licenses for standalone computers.

Integrated Hardware and Software – a Joint Commitment to Efficiency

Dedicated to getting each customer up and running with the integrated SEQUEST Cluster system as soon as possible, Thermo Electron will work together with one of three supported computer hardware vendors, chosen by the customer, to deliver the entire solution fully racked, cabled, installed with operating system and application software, and tested onsite, ready for checkout by a joint team of engineers from each company. This commitment to cooperation ensures efficiency during the installation and configuration process, and extends the value of the solution by minimizing the need for support or service later.

The SEQUEST Cluster will be supported on specific, customized configurations from the following three vendors. IBM® has been chosen as the preferred vendor for this integrated solution for their commitment to product quality, value, and for their network of world-wide support.

- **IBM**
- **Dell™**
- **RackSaver®**

Please contact your local Thermo Electron sales representative for information on supported hardware systems for the SEQUEST Cluster software and how to order the specified configurations, or visit our website at analyze@thermo.com to place a request for more information.

SEQUEST Cluster Hardware: Recommended Configuration Options

The following list includes the recommended configuration of the management and computing nodes for the systems supplied by the three preferred computer vendors.

Processors	Intel Xeon™ - 2.4 GHz
L2 Cache	512 KB
RAM Memory	512 MB to 4 GB for system expansion
SCSI Internal Storage	18.2 GB to 440.4 GB
IDE Internal Storage	40 GB to 240 GB
System Connectivity	One 10/100 Mbps Ethernet switch standard
Adapters	Dual Gigabit Ethernet SX, Myrinet 133 MHz, 10/100 Mbps Ethernet, Server RAID-4LX, Ultra320 SCSI
System Control	Minimum of one Keyboard/video/Mouse (KVM) switch required
Miscellaneous Options	Floppy Disk Drives High-Speed Interconnect Networking Support Line and Warranty Services

The Comprehensive Package - BioWorks 3.1

The SEQUEST Cluster package includes all of the innovative utilities within the BioWorks 3.1 Protein Identification Software Suite that have made it the most comprehensive protein identification software suite available:

- **XPRESS for relative quantitation**
- **ZSA for accurate charge state analysis**
- **Multiple File Consensus to compare and consolidate search results from different samples**
- **PEPMATCH™ and PEPMAP™ for annotating results**
- **Database Manager for customized database indexing and retrieval**
- **Optional license for SALSA pattern recognition algorithm**

Accurate, Specific, and Faster than Ever

At the core of the Cluster is the SEQUEST protein identification algorithm, which uses the specificity of MS/MS spectra to identify proteins by matching the spectra of digested peptides with computer-generated fragmentation patterns predicted from protein sequence databases. Identifying even low abundant proteins from complicated mixtures such as those from whole cell lysates and human tissue samples, SEQUEST has become an essential tool during data interpretation for demanding proteomics applications.

Complementing Protein I.D. with Powerful Data Mining Tools

- Automating relative quantitation is often considered the first step towards protein expression analysis by measuring discreet changes in the protein levels in cells. **XPRESS** can quantify peptides tagged with any label, including cleavable and non-cleavable ICAT™ reagent-labeled peptides. Enhancements to the utility now allow quantifying relative changes in peptides labeled on their N- or C-terminus.

- **ZSA** is an algorithm that more accurately determines the charge state of a precursor ion in order to reduce the number of spectra submitted for a SEQUEST search, reducing the number of +2 and +3 peptide duplicates that are both searched, and ultimately improving productivity.

- The **Multiple File Consensus** report consolidates and sorts the results from as many as 12 different SEQUEST searches, simplifying multi-dimensional protein identification, and enabling comparisons of peptides and post-translational modifications from analogous proteins.

- **PEPMATCH** and **PEPMAP** are utilities used to customize and annotate spectral plots, as well as to determine where specific peptides may have eluted within the chromatogram.

- Protein databases can be retrieved on a regular basis from public websites, indexed to increase their specificity and efficiency, and customized with proprietary proteins using the **Database Manager** tool.

- **SALSA** is a spectral pattern recognition tool that mines MS/MS spectra for sequence motifs in targeted protein studies to search for unexpected modifications or mutations, and can be included to the SEQUEST Cluster/BioWorks 3.1 package as an additional option.

In addition to these offices, Thermo Electron Corporation maintains a network of representative organizations throughout the world.

Australia
+61 2 9898 1244 • analyze.au@thermo.com

Austria
+43 1 333 50340 • analyze.at@thermo.com

Belgium
+32 2 482 30 30 • analyze.be@thermo.com

Canada
+1 800 532 4752 • analyze.ca@thermo.com

China
+86 10 5850 3588 • analyze.cn@thermo.com

France
+33 1 60 92 48 00 • analyze.fr@thermo.com

Germany
+49 6103 4080 • analyze.de@thermo.com

Italy
+39 02 950 591 • analyze.it@thermo.com

Japan
+81 45 453 9100 • analyze.jp@thermo.com

Latin America
+1 512 251 1503 • analyze.la@thermo.com

Netherlands
+31 76 587 98 88 • analyze.nl@thermo.com

Nordic
+46 8 556 468 00 • analyze.se@thermo.com

South Africa
+27 11 570 1840 • analyze.sa@thermo.com

Spain
+34 91 657 4930 • analyze.es@thermo.com

Switzerland
+41 61 48784 00 • analyze.ch@thermo.com

UK
+44 1442 233555 • analyze.uk@thermo.com

USA
+1 800 532 4752 • analyze.us@thermo.com

www.thermo.com

ISO 9001 REGISTERED COMPANY

Thermo Finnigan LLC, San Jose, CA USA is ISO Certified.

BR61371_E 12/03S

Thermo
ELECTRON CORPORATION