

What is Sequence Alignment ?

Genome sequencers produce large numbers of relatively short (from tens to hundreds of base-pair) sequences, which come from randomly sampled, unknown locations along the genome being sequenced. Without knowing the location from which each sequence originated, it has no practical use. The process of mapping each of these short sequences to unique chromosome and position within that chromosome is known as *Sequence Alignment*. Sequence Alignment is done by comparing each sequence with a database (known as a *reference*) of a complete genome of the same species. The comparison is done with an allowance for some discrepancies to account for machine errors (known as *sequencing errors*) as well as actual variations between different individuals of the same species (*genomic variation*). Under the ideal assumptions that the genome is not repetitive, and that the sequencing error as well as the genomic variation rates are low, alignment should be able to correctly map all the sequences coming out of the sequencer. Realistically, using the human genome as an example, it is believed that approximately 75 percent of the sequences, which the sequencing machine produces, can be uniquely and accurately mapped by proper sequence alignment, while the remaining sequences are discarded due to mapping uncertainty. This rate may be improved using “mate pair” techniques. For less repetitive genomes, such as those of micro-organisms with compact genomes, higher mapping rates are typically achieved.

```
...TCAAGATTGAAATATAGGGCGCTCGGGAGAGAGATTCACCATCGCTAACGATACCGGTCCAATG...  
      AATATAGGGCGCTCGGGAGAGAAAT
```

Figure 1: Example of Sequence Alignment with One Mismatch

The sequence alignment process can be extremely computationally intensive, but is a critical step in genome sequencing, since no further processing can be done on the sequences detected by the sequencer, until their location within the genome is deciphered.

What Does Sequence Alignment Require ?

When larger organisms (e.g. human) are sequenced with coverage ratios adequate for statistically meaningful results, scientists are confronted with the reality of computationally intensive sequence alignment. Using standard computers, they may have to wait weeks after running a sequencing job, to get any useable data. For some, the solution is to spend more money on large amounts of computers, known as *clusters*, or *compute farms*. These clusters can consume large amounts of electricity, require real estate with very expensive air-conditioning to keep them from damage due to overheating, and usually require full-time IT professionals to operate. This is why industry analysts now predict that for every Dollar spent on Next Generation Sequencing (NGS) equipment, *another* Dollar is spent on IT infrastructure to process the raw results. While, for example, the Genome Sequencing Center at Washington University in St. Louis benefits from an \$11 million, 16,000 square foot IT facility

specifically designed to handle the computational requirements of their NGS machines, most of the scientists who need NGS for their research are outside the large genome centers. Some of these scientists find it increasingly challenging to make full use of the high throughput of the new sequencing machines. Thanks to the Imagenix Sequence Alignment System (ISAS), these scientists can perform their leading edge research *without* the million Dollar compute farms.

What is ISAS ?

ISAS can be run on a single computer, and produce the same results, in less time than the standard sequence alignment software does when running on a big cluster of computers.

During a recent demonstration performed at Applied Biosystems in Foster City, California, **ISAS (version 32 with native colorspace) completed the mapping of 100 million individual 25mers (sequences of 25 base pairs) with an allowance for two substitutions, against the 3 Billion base human genome in just 37 minutes**. The ISAS software was running on a single computer which is similar (but slightly slower) to the ones Applied Biosystems ships with their SOLiD machine. The input file was taken from a SOLiD machine after sequencing a human DNA sample. The same task, using the standard sequence alignment tool that comes with SOLiD, took over 18 hours on the Applied Biosystems compute farm cluster. The results were identical, and written in the same file format. Note that this demonstration is approximately “1x” coverage for a human genome, We expect realistic sequencing to be done at “5x” or higher coverage, generating 5 times more sequences to be aligned.

Using a cluster of computers, the run time decreases approximately in a linear rate, so on a cluster of 10 blades, 1 billion 25mers can be aligned in under 40 minutes (from the time the input file is written on the computers hard disk).

How Does ISAS Work ?

Imagenix engineers and scientists combined skills from the life sciences with skills from computer engineering to create what we consider the world’s fastest sequence alignment systems.

Each ISAS system actually includes two different sequence alignment algorithms. For each sequence to be mapped, the system first does a simple analysis in order to determine, based on the known statistics of the reference, which of the two algorithms is to be used for optimal speed on each sequence.

In 1971 a small company by the name of Intel publicly introduced the world's first single chip microprocessor, the Intel 4004. The architecture of this ingenious little chip, which transformed computing forever, was designed by Intel engineer Ted Hoff. Ted had just completed his Ph.D. under the supervision of professor Bernard Widrow at the nearby Stanford University. ISAS architecture was designed by Hadar Isaac, who also

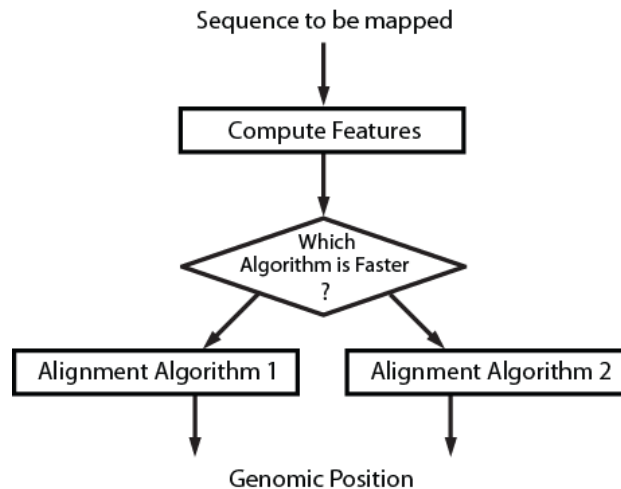


Figure 2: Dual Alignment Algorithms in ISAS

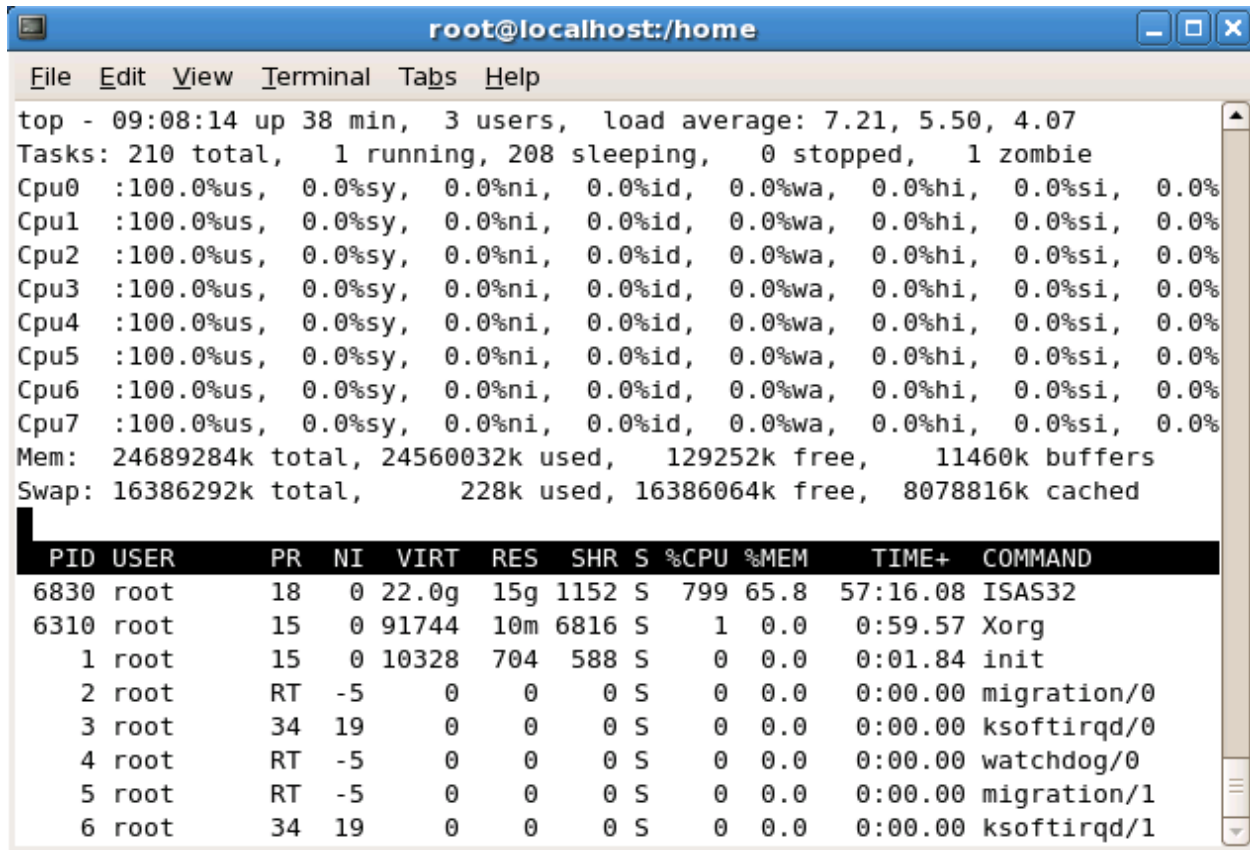
completed his Ph.D at the Information Systems Lab of Stanford University, under the supervision of the legendary Professor Widrow. After over three decades of evolution, the processors that provide computational power to today's computers have mushroomed into amazingly powerful technological wonders. But with increasing processor power comes increasing processor complexity, recently compounded by the latest processor technology breakthrough: multi-core CPUs.

ISAS software makes use of direct machine instruction commands, often bypassing the inefficiencies of compiled language implementations. This technique is rarely used by other software applications. In addition, ISAS employs advanced memory-management techniques to optimize the use of the fast cache memories which are included with today's advanced processors. For example, the processor chips found in the SOLiD

machine have 12Mbytes of level 2 cache memory inside each CPU chip. This memory is several times faster than the computer's main RAM memory. As more of this 12MB of ultra fast memory is deployed by a computer program, the faster this program will inevitably run.

Finally, ISAS makes use of Imagenix proprietary load balancing technology to keep all the CPU cores in the computer running at as close to 100% of their rated performance as possible, essentially squeezing every drop of computer power from today's amazingly powerful microprocessor chips. Typical computer applications only make use of a small percentage of the available computing available, usually because few engineers have the skills to exploit them properly. Figure 3 below shows a real-time snapshot of ISAS CPU utilization on an 8 CPU core (2 quad-core) system. ISAS is utilizing 7.99 CPU cores out of 8.00.

ISAS is available in multiple configurations. Although the simplest configuration uses one computer, and still outperforms computer clusters, even further performance gains can be achieved by unleashing ISAS on multiple computers, scaling almost linearly with the number of computers.



```
root@localhost:/home
File Edit View Terminal Tabs Help
top - 09:08:14 up 38 min, 3 users, load average: 7.21, 5.50, 4.07
Tasks: 210 total, 1 running, 208 sleeping, 0 stopped, 1 zombie
Cpu0 :100.0%us, 0.0%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%
Cpu1 :100.0%us, 0.0%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%
Cpu2 :100.0%us, 0.0%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%
Cpu3 :100.0%us, 0.0%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%
Cpu4 :100.0%us, 0.0%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%
Cpu5 :100.0%us, 0.0%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%
Cpu6 :100.0%us, 0.0%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%
Cpu7 :100.0%us, 0.0%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%
Mem: 24689284k total, 24560032k used, 129252k free, 11460k buffers
Swap: 16386292k total, 228k used, 16386064k free, 8078816k cached

  PID USER      PR  NI  VIRT  RES  SHR  S  %CPU  %MEM    TIME+  COMMAND
 6830 root        18   0 22.0g 15g 1152  S   799  65.8   57:16.08 ISAS32
 6310 root        15   0 91744 10m 6816  S    1   0.0    0:59.57 Xorg
    1 root        15   0 10328  704  588  S    0   0.0    0:01.84 init
    2 root         RT  -5    0    0    0  S    0   0.0    0:00.00 migration/0
    3 root        34  19    0    0    0  S    0   0.0    0:00.00 ksoftirqd/0
    4 root         RT  -5    0    0    0  S    0   0.0    0:00.00 watchdog/0
    5 root         RT  -5    0    0    0  S    0   0.0    0:00.00 migration/1
    6 root        34  19    0    0    0  S    0   0.0    0:00.00 ksoftirqd/1
```

Figure 3: Maximally Efficient CPU Utilization in ISAS

ISAS is available as software to run on the computers included with your SOLiD system, or on your own computer (if it meets the minimum requirements). It is also available pre-installed as a turn key system on an Imagenix computer. Imagenix computers are optimized for running ISAS, as well as most genomics applications, faster than off-the-shelf computers.

ISAS is the fastest alignment system known. [Performance benchmarks](#) can be viewed on our web site.

How Can I Learn More about ISAS ?

For more information, contact us or visit our web site.

Imagenix Technologies
171 Main Street #108
Los Altos, CA 94022

Tel. (650) 917-9998
Fax (650) 917-8765

www.imagenix.com/genomics