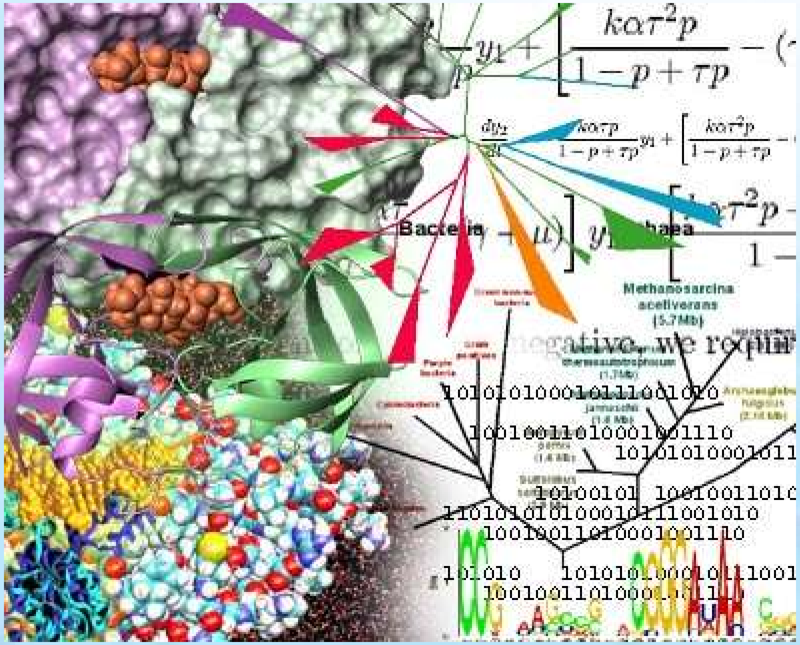
A Grid implementation of the sliding window algorithm for protein similarity searches facilitates whole proteome analysis on continuously updated databases

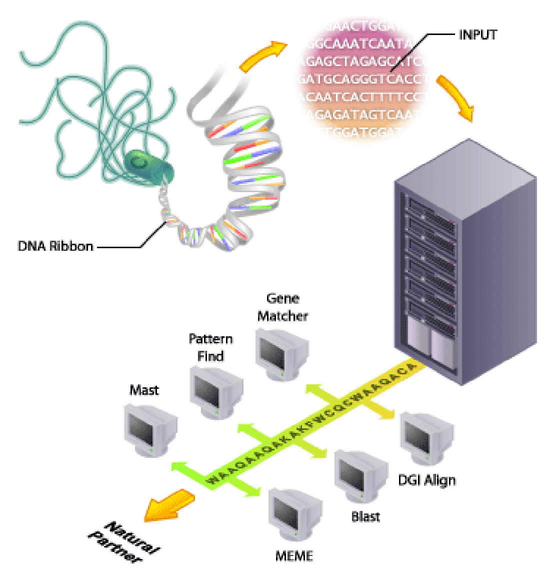
Jorge Andrade
Department of Biotechnology, Royal Institute of Technology (KTH), Stockholm, Sweden.

Bionformatics



Bioinformatics involves the integration of computers, software tools, and databases in an effort to address biological questions

BLAST

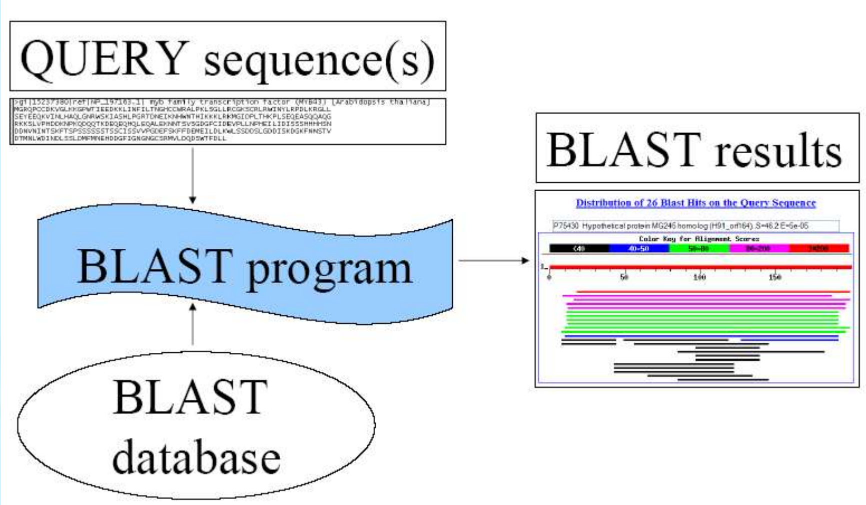


The Blast algorithm

The **BLAST** programs (**B**asic **L**ocal **A**lignment **S**earch **T**ools) are a set of sequence comparison algorithms introduced in 1990 that are used to search sequence databases for optimal local alignments to a query.

Manual alignment

```
           GATGCCATAGAGCTGTAGTCGTACCCT   <–      Seq. A
Seq. B –>    CTAGAGAGC-GTAGTCAGAGTGTCTTTGAGTTCC
```



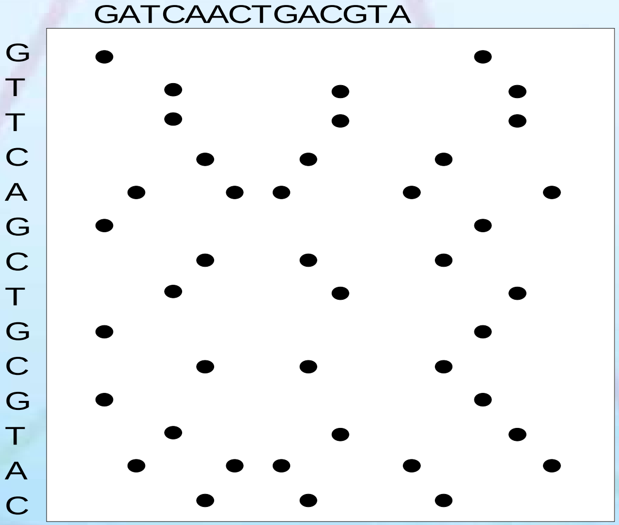
QUERY sequence(s)

BLAST program

BLAST database

BLAST results

Distribution of 26 Blast Hits on the Query Sequence



Simple Dot Plot

Alignment scores: match vs. mismatch

Simple scoring scheme (too simple in fact…):

Matching amino acids:	5
Mismatch:	0

Scoring example:

```
K A W S A D V
:   : : :   :
K D W S A E V
5+0+5+5+5+0+5 = 25
```

Protein substitution matrices

A	5																			
R	-2	7																		
N	-1	-1	7																	
D	-2	-2	2	**8**																
C	-1	-4	-2	-4	13															
Q	-1	1	0	0	-3	7														
E	-1	0	0	**2**	-3	2	**6**													
G	0	-3	0	-1	-3	-2	-3	8												
H	-2	0	1	-1	-3	1	0	-2	10											
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5										
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5									
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6								
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7							
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8						
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10					
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	**5**				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	**2**	**5**			
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15		
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

BLOSUM50 matrix:

- Positive scores on diagonal (identities)

- Similar residues get higher scores

- Dissimilar residues get smaller (negative) scores

Pairwise alignments

```
43.2% identity;                   Global alignment score: 374

                10        20        30        40              50
alpha  V-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSA
       : :.: .:. : : ::::  .. : :.::: :... .: :. .:  : :::      :.
beta   VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
               10          20        30        40        50

            60        70        80        90       100       110
alpha  QVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL
       .::.:::::  :.....::.:.. .....::.::  ::.::: ::.::.. :. .:: :.
beta   KVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHF
       60        70        80        90       100       110

           120       130       140
alpha  PAEFTPAVHASLDKFLASVSTVLTSKYR
         :::: :.:. .: .:.:...:. ::.
beta   GKEFTPPVQAAYQKVVAGVANALAHKYH
      120       130       140
```

Why Compare Sequences?

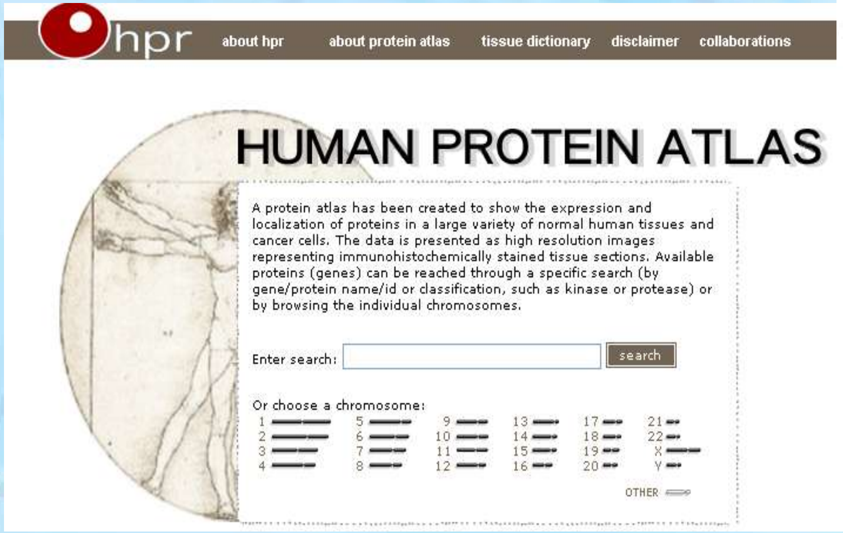
What biologists do with blastp?

-Predicting a protein function

-Predicting a protein 3-D structure

-Finding protein family members

-Antibody recognition site



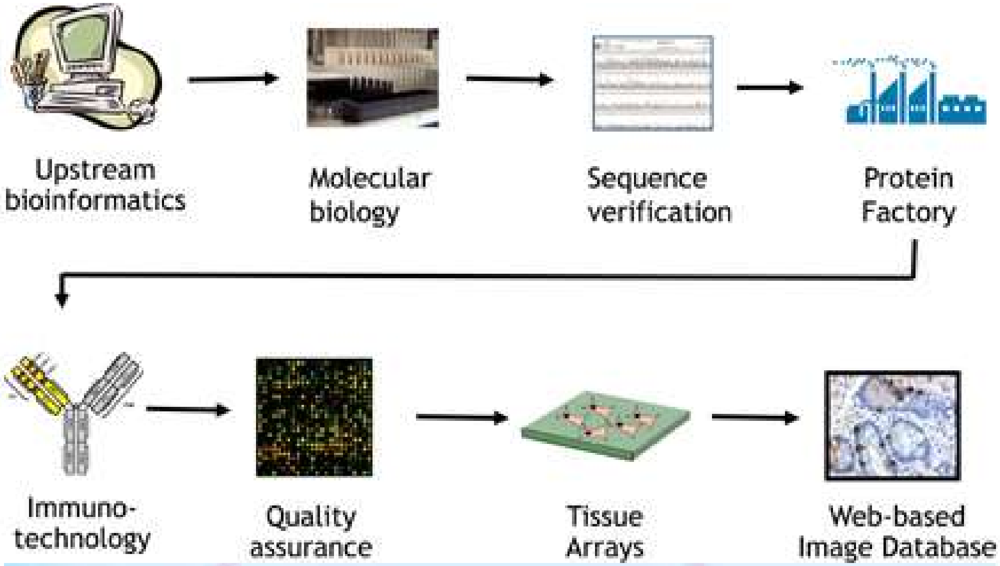
HUMAN PROTEIN ATLAS

A protein atlas has been created to show the expression and localization of proteins in a large variety of normal human tissues and cancer cells. The data is presented as high resolution images representing immunohistochemically stained tissue sections. Available proteins (genes) can be reached through a specific search (by gene/protein name/id or classification, such as kinase or protease) or by browsing the individual chromosomes.

Enter search:

Or choose a chromosome:

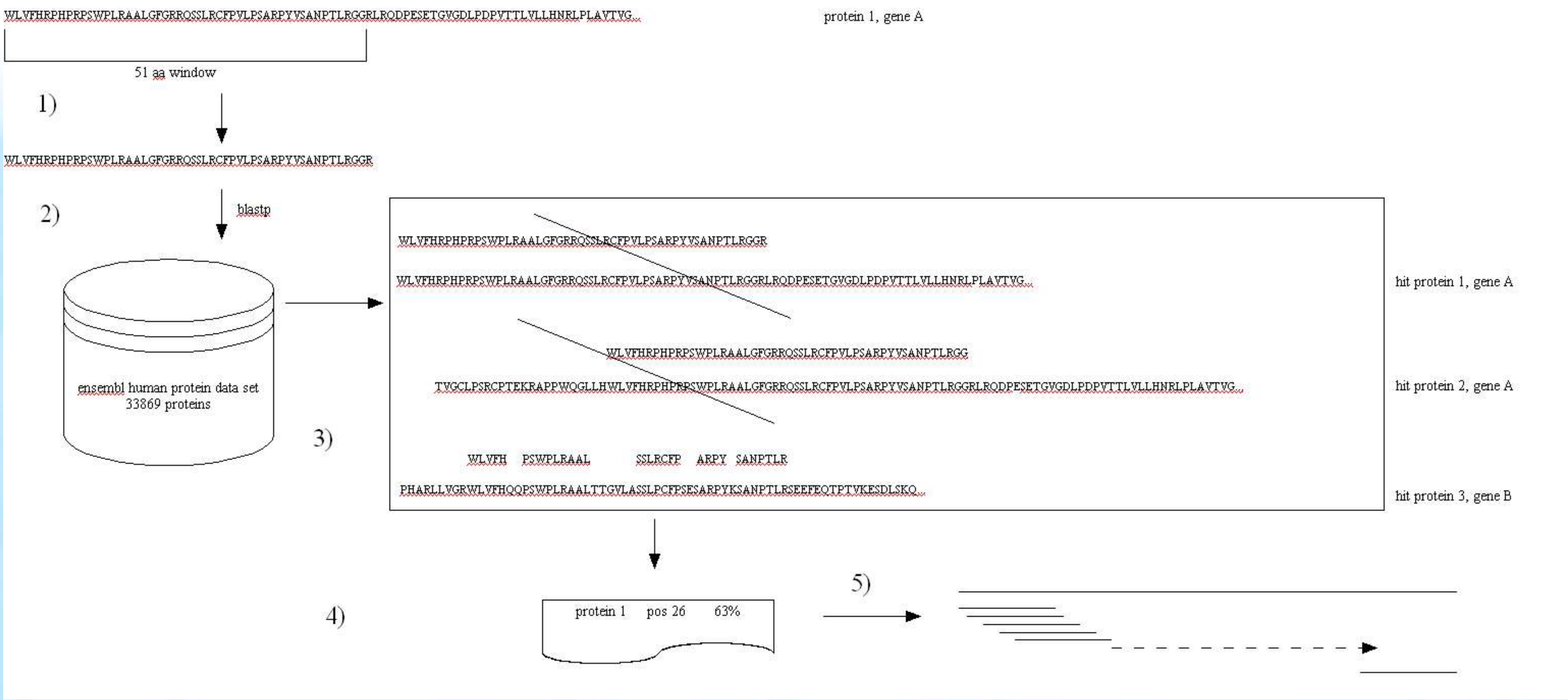
www.hpr.se



Upstream bioinformatics → Molecular biology → Sequence verification → Protein Factory → Immuno-technology → Quality assurance → Tissue Arrays → Web-based Image Database

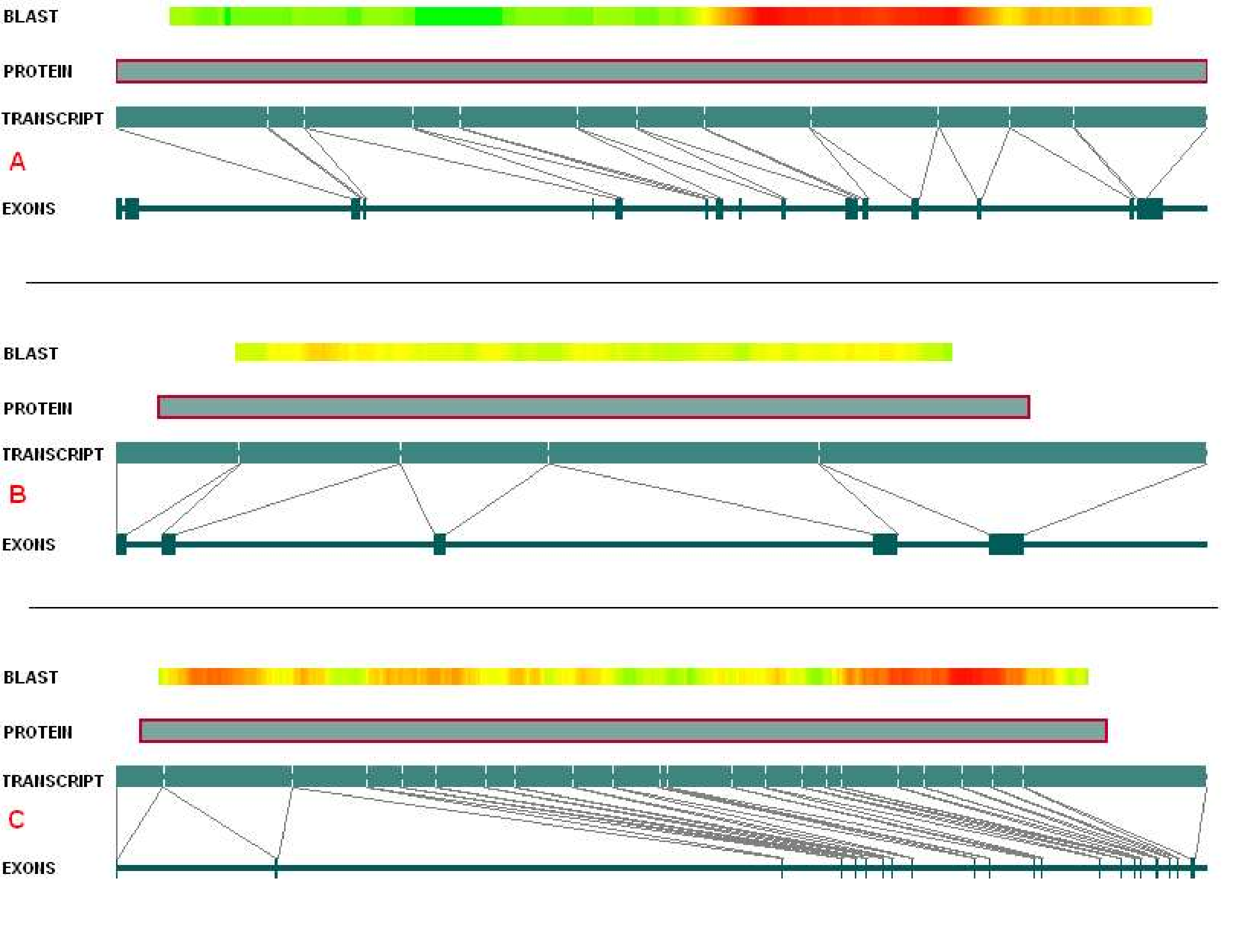
- Select a unique fragmet of a protein
- Express that protein fragment in laboratory
- Immunice protein fragment to rabbit
- Rabbit create the anti bodies
- PrEST
- Validation of antibodies (no crossbinding)
- Color label atibodies
- Antibody on differen tissues, binding to protein.

Sliding window protein similarity search


WLVFHRPHPRPSWPLRAALGFGRRQSSLRCFPVLPSARPYVSANPTLRGGRLRQDPESETGVGDLPDPVTTLVLLHNRLPLAVTVG...
protein 1, gene A
51 aa window
1)
WLVFHRPHPRPSWPLRAALGFGRRQSSLRCFPVLPSARPYVSANPTLRGGR
2)
blastp
ensembl human protein data set
33869 proteins
3)
WLVFHRPHPRPSWPLRAALGFGRRQSSLRCFPVLPSARPYVSANPTLRGGR
WLVFHRPHPRPSWPLRAALGFGRRQSSLRCFPVLPSARPYVSANPTLRGGRLRQDPESETGVGDLPDPVTTLVLLHNRLPLAVTVG...
hit protein 1, gene A
WLVFHRPHPRPSWPLRAALGFGRRQSSLRCFPVLPSARPYVSANPTLRGG
TVGCLPSRCPTEKRAPPWQGLLHWLVFHRPHPRPSWPLRAALGFGRRQSSLRCFPVLPSARPYVSANPTLRGGRLRQDPESETGVGDLPDPVTTLVLLHNRLPLAVTVG...
hit protein 2, gene A
WLVFH PSWPLRAAL SSLRCFP ARPY SANPTLR
PHARLLVGRWLVFHQQPSWPLRAALTTGVLASSLPCFPSESARPYKSANPTLRSEEFEQTPTVKESDLSKQ...
hit protein 3, gene B
4)
protein 1 pos 26 63%
5)

The protein fragments, denoted Protein Epitope Signature Tags (PrESTs), comprise 100 to 150 amino acids (2). PrEST design is based on the selection of a protein region with as low as possible similarity to protein regions from other genes. This is important to avoid cross-reactivity of the resulting antibody.

Graphical representation



Graphical representation where the identity of a 51 amino acid fragment of the target protein to all other human proteins from other genes is displayed as a color_coded line at the middle position of the fragment on the protein. Green color code implies <40% identity, yellow 40-60%, orange >60-80% and red >80% identity

The problem

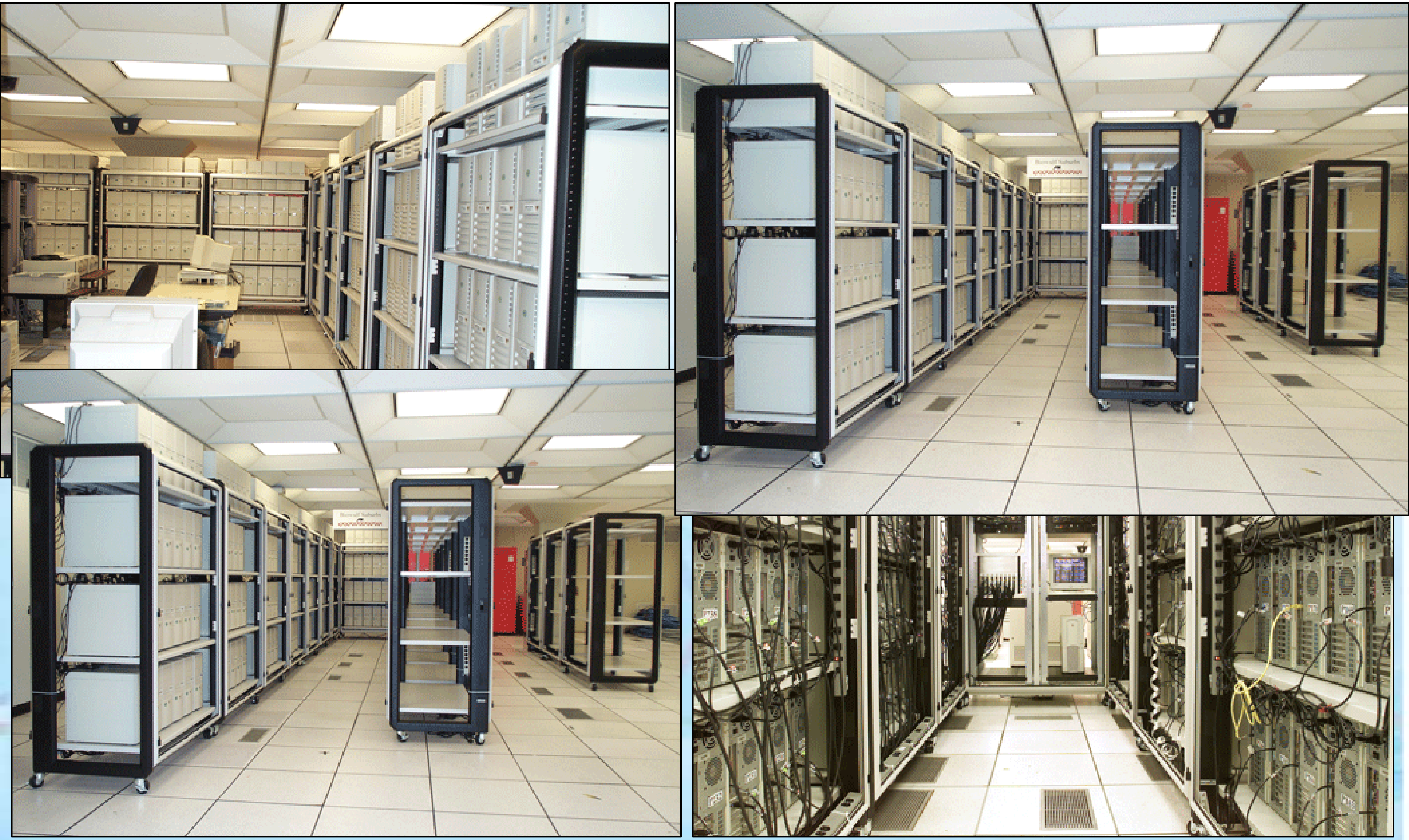
When using the complete Ensembl human protein data set (version 31.35) with 33869 sequences as input, the runtime on a single up-to-date workstation is 1300 hours. This task comprises a total of 15,193,041 blastp searches

15,193,041 blastp searches

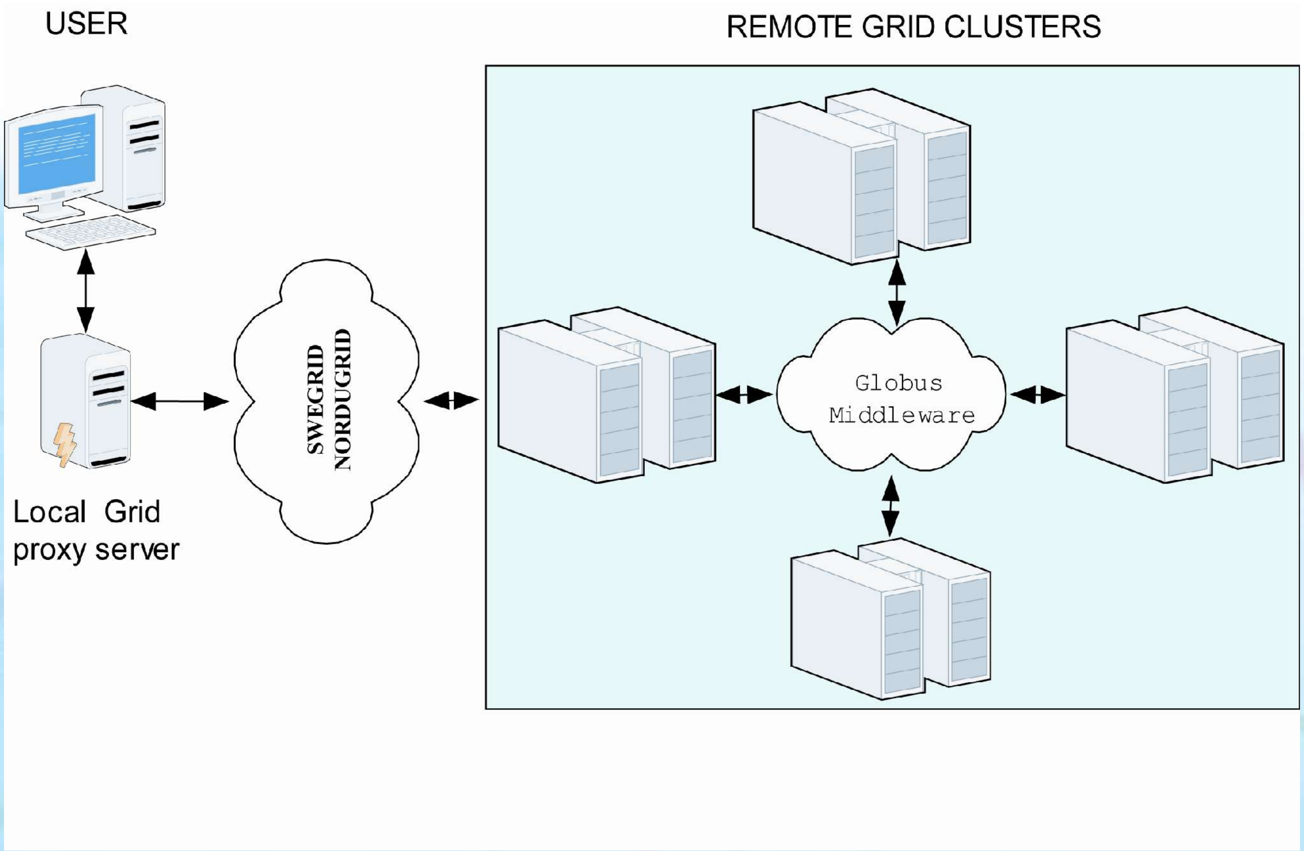
8 weeks

Ensembl is a continuously updated database, generally once a moth.

The solution

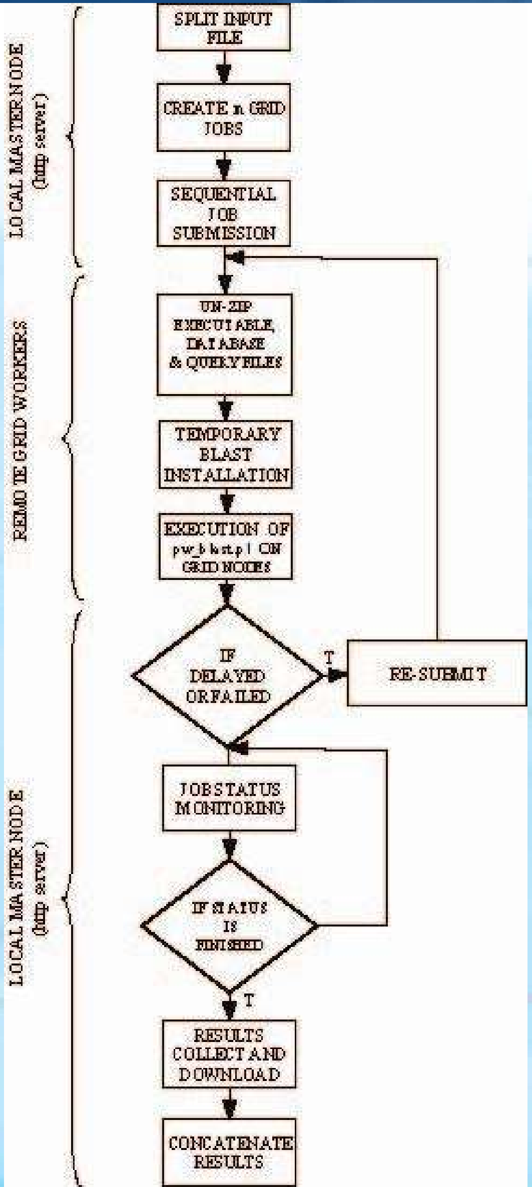


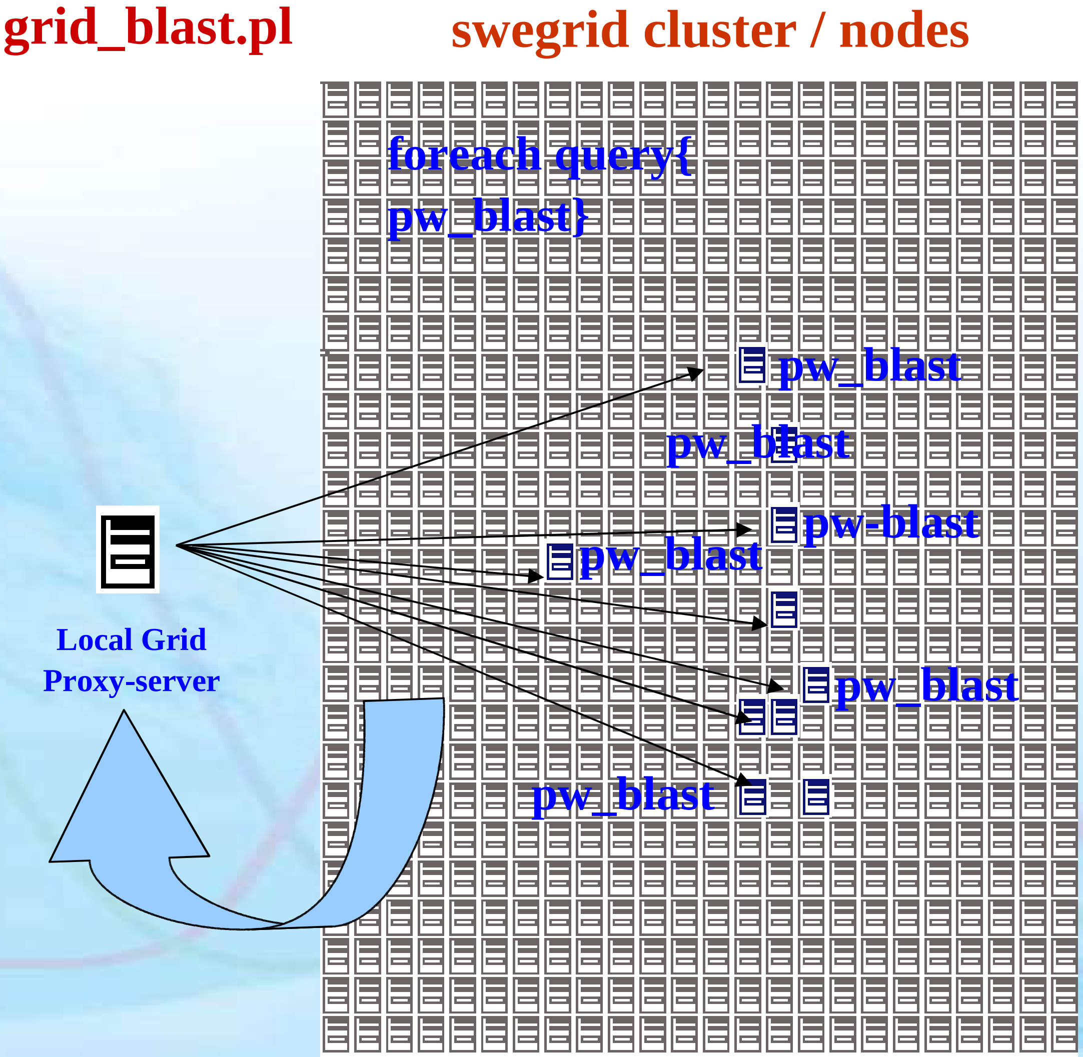
Grid – Blast Architecture



To develop and implement this in a Grid environment, we joined the Swegrid / NorduGrid virtual organization. We were granted by Swedish National Infrastructure for Computing (SNIC) to have access to ~600 nodes, 1000 h/month through the different Swedish clusters.

Grid broker


LOCAL MASTER NODE (http server)
SPLIT INPUT FILE
CREATE n GRID JOBS
SEQUENTIAL JOB SUBMISSION
REMOTE GRID WORKERS
UN-ZIP EXECUTABLE, DATABASE & QUERY FILES
TEMPORARY BLAST INSTALLATION
EXECUTION OF pw_blast.pl ON GRID NODES
IF DELAYED OR FAILED
T
RE-SUBMIT
LOCAL MASTER NODE (http server)
JOBSTATUS MONITORING
IF STATUS IS FINISHED
T
RESULTS COLLECT AND DOWNLOAD
CONCATENATE RESULTS



grid_blast.pl

swegrid cluster / nodes

foreach query{
pw_blast}

pw_blast

pw_blast

pw-blast

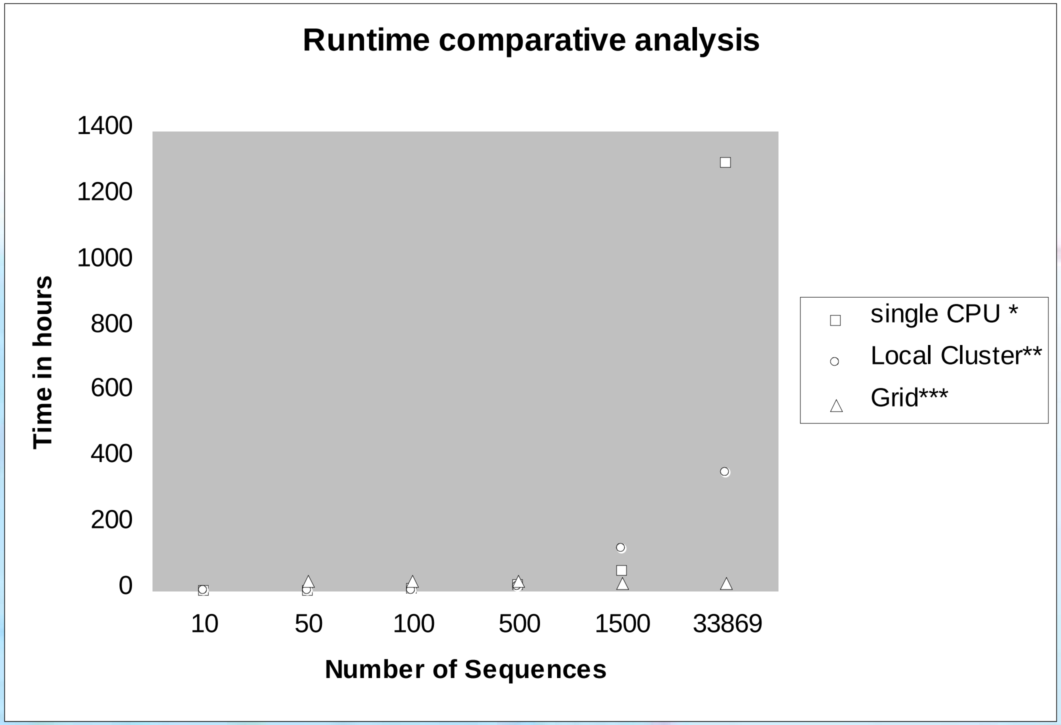
pw_blast

pw_blast

pw_blast

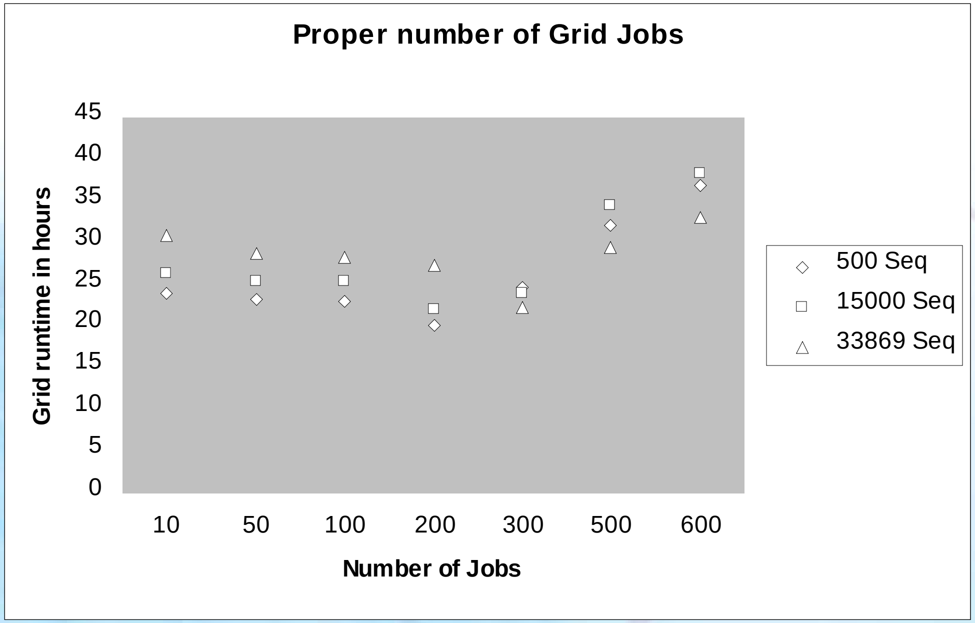
Local Grid
Proxy-server

Results



*Runtime comparative analysis: *single CPU 1Ghz speed/512Mb RAM, ** local cluster with 5 processor units each 1Ghz speed/512Mb RAM, *** Swegrid environment with access to ~600 remote CPUs with similar or better hardware. The Grid analysis was performed by submitting the sequence in file split into 300 atomistic jobs. The runtime for the analysis of the complete Ensembl human protein data (33869 protein sequences) was reduced from* ***1304 hours on a single CPU to 22,3 hours on the Grid****. The analysis has been repeated several times. The exact Grid runtime can vary depending to different Grid conditions but the overall performance relative to a single CPU is marginally affected.*

Proper number of Grid Jobs



Proper number of Grid jobs. The chart shows the runtime needed for three different size input data sizes, 500, 15000 and 33869 sequences long input files. The time needed to submit the complete set of jobs to the Grid nodes has to be approximate the same as the time needed for a single node to run a single atomistic part of the complete set of jobs

CONCLUSION

- If the time for submitting the complete set of jobs to the Grid exceeds the time to execute a single atomistic job, the data input has been sub-optimally split into.

- Grid implementations for computer intensive Bioinformatics tasks represents an economical and time efficient alternative.

- A local TEMPORARY installation of the executable and database upon submission, makes it very suitable for dynamic environments, avoids the need for a predefined environment , and does not leave/take up space on the computer between runs.