SEQR: searching protein sequences by similarity in Solr

Lewis Geer

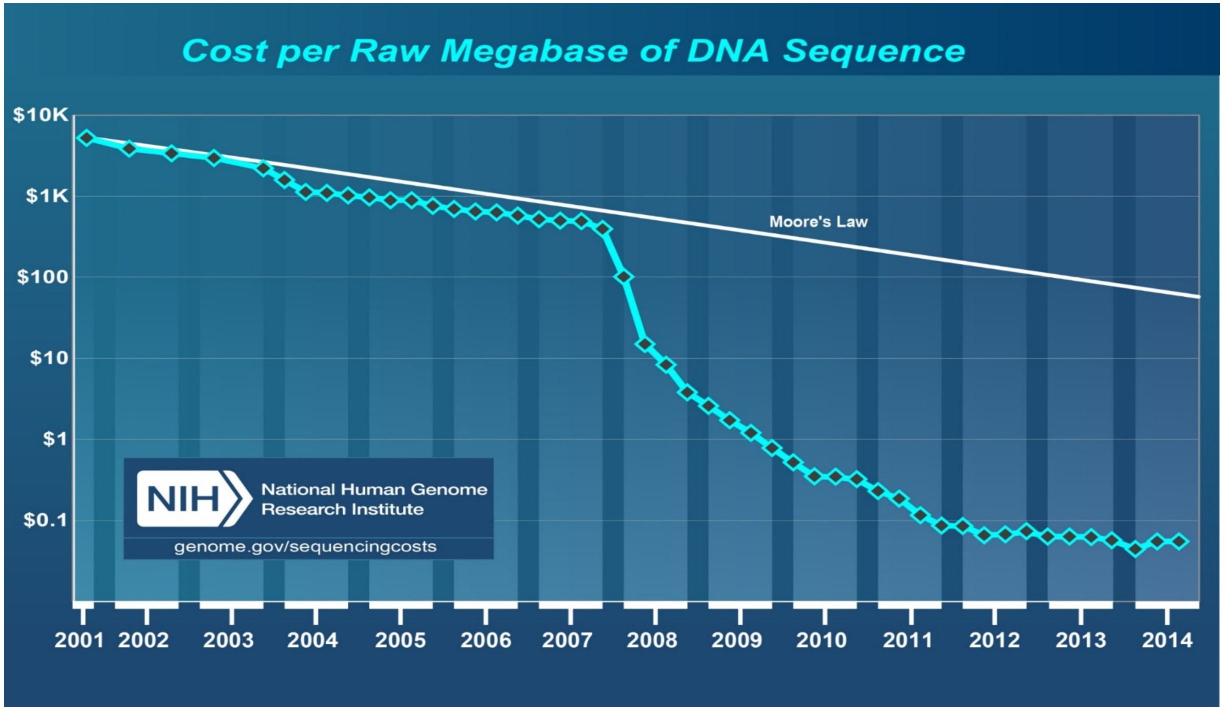
National Institutes of Health

NCBI

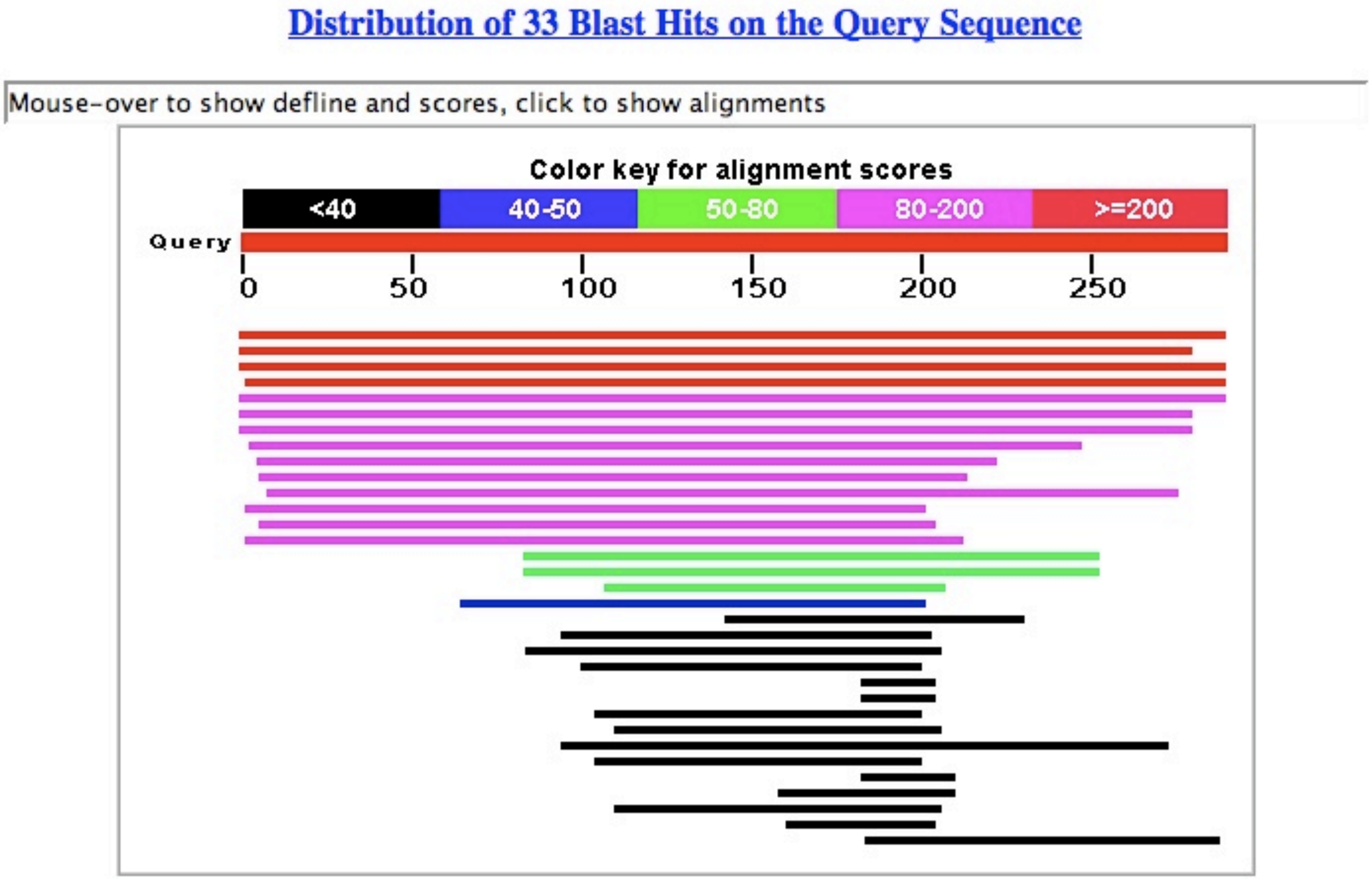
Questions

- Algorithm
  - – Can we create a similarity search that is computationally efficient? Space efficient?
- Implementation
  - – Can we improve the usability of sequence search in the era of nextgen sequencing?
  - – Can we implement the algorithm by leveraging open source to minimize development costs?

Similarity indexes become necessary as sequencing beats Moore's law



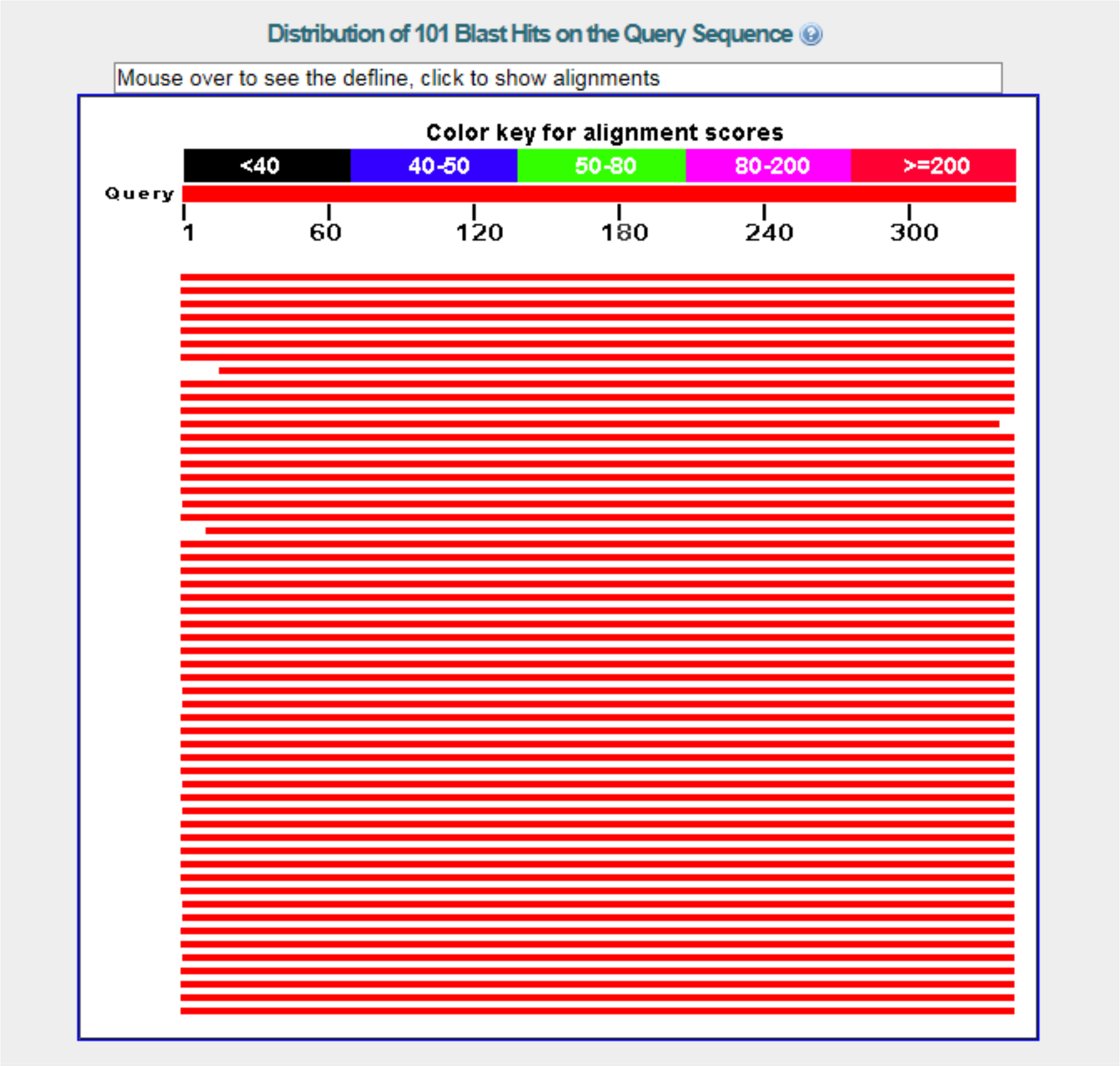
BLAST 2006



Distribution of 33 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

BLAST 2015



Distribution of 101 Blast Hits on the Query Sequence

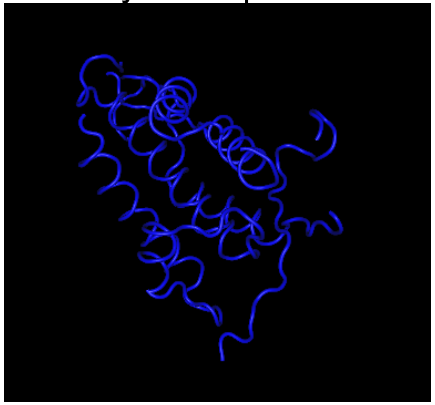
Mouse over to see the defline, click to show alignments

How do we provide useful results quickly?

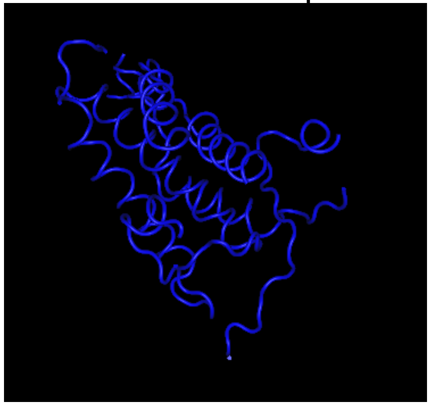
- Surface interesting sequences first
- Integrate queries into search results
  - Allow useful sub-setting after the search, e.g. taxonomy
  - Faceting
  - Autosuggestion
- Display information in more understandable ways, e.g. trees.
- Quick response to allow iterative discovery
- Reconfigure interface to fit specific user needs

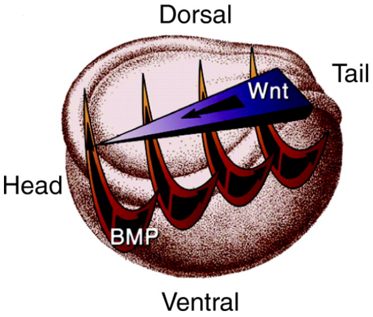
Demo

Fly Wnt8 protein

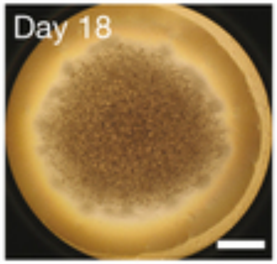


Human Wnt8 protein



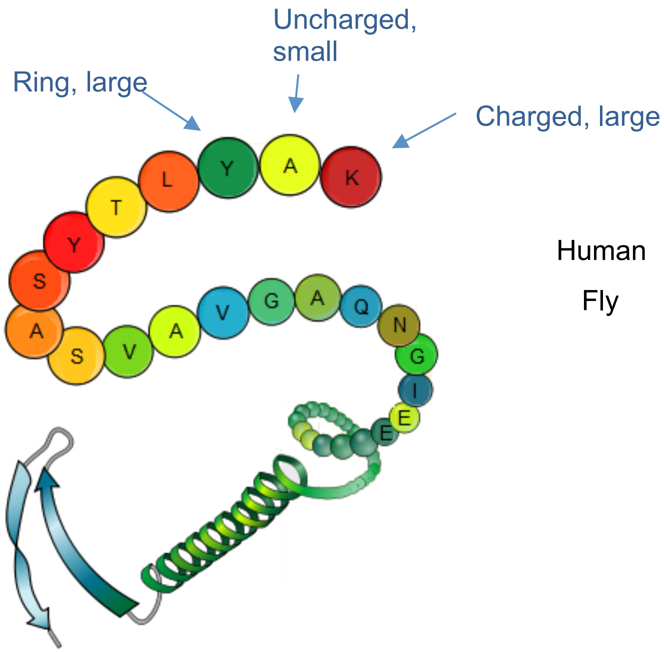


Fly embryo layout



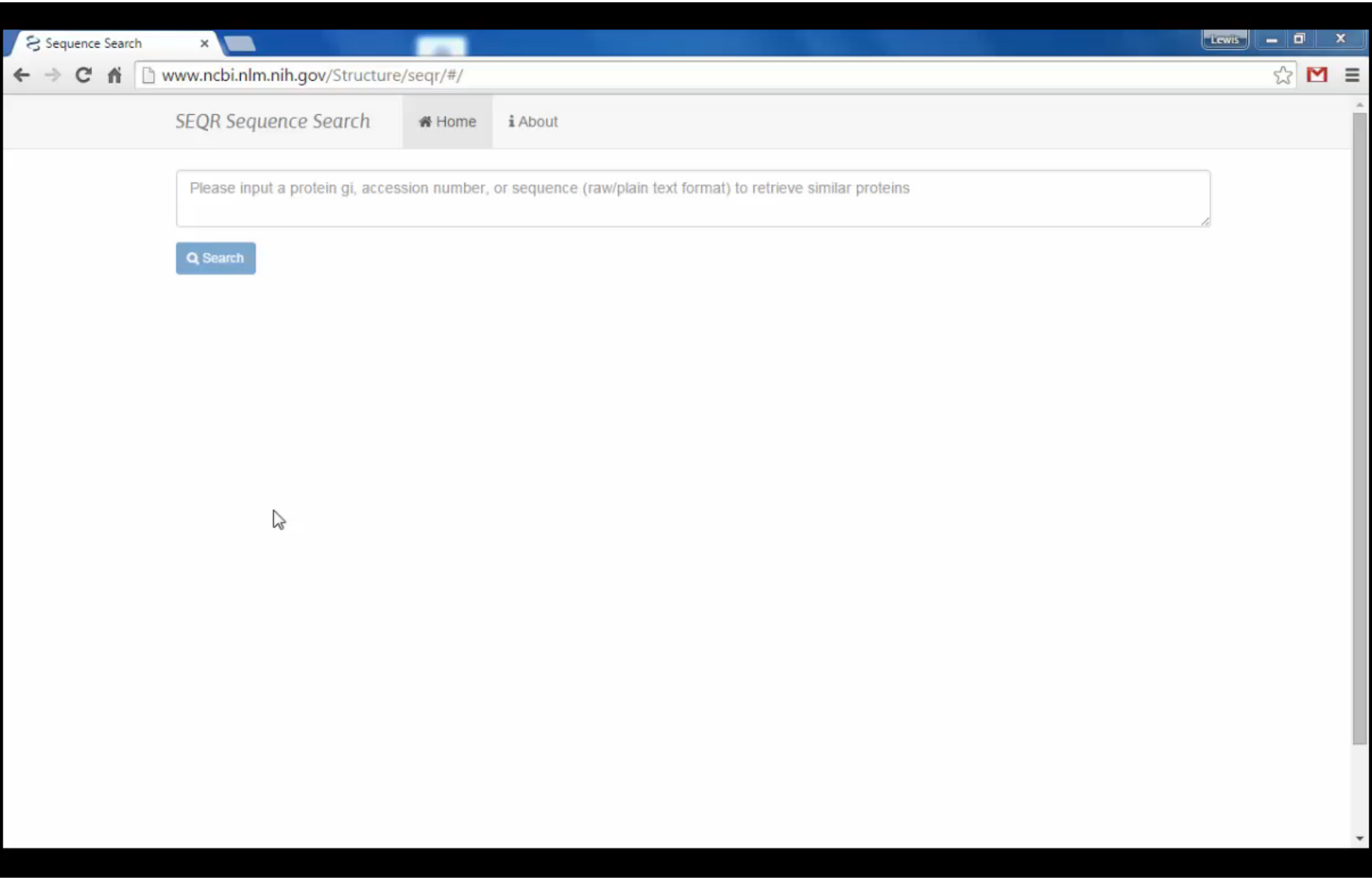
Human kidney
Grown in lab

Melissa Little et al.



```
Human  KAYLTYSASVAVGAQNGIEECKYQFAWERWNCP-ESTLQLATHNGLRSATRETSFVHAIS
       +A L++      G +  ++ C+  F W+RWNCP +  +Q  +     S  RE  +V AIS
Fly    QAPLSWEDITGKGLKQALDSCQQSFQWQRWNCPSQDFVQKNSKPEENSPNREDVYVAAIS

       SAGVMYTLTRNCSMGDFDNCGCDDSRNGRIGGRGWVWGGCSDNAEFGERISKLFVDGLET
        A +++TLT++C+ G    CGC ++                +  +  E+  K F
       MAAIVHTLTKDCANGVIAGCGCTENALNVPCAH--------EPTKALEQYEKHF------

       GQDARALMNLHNNEAGRLAVKETMKRTCKCH---GISGSCSIQTCWLQLAEFRDIGNHLK
       G  + A+   HN       ++ ++++ C+C     + G C  + C   L  F  I   L
       GSGSGAIG--HNRRVVGALLQRSLEQECRCKQPGAVQGECQEEECVAVLKPFEAIAQDLL

       IKHDQALKLEMDKRKMRSGNSADNRGAIADAFSSVAGSELIFLEDSPDYCLKNISLGLQG
         +D A++LE        G S++    +   + ++    L+F++DSP+YC ++ + GL G
       QMYDDAIQLE--------GASSN----LKIMWQNIPLDSLVFMQDSPNYCERDAT-GLWG
```

http://www.ncbi.nlm.nih.gov/Structure/seqr/



Sequence Search
www.ncbi.nlm.nih.gov/Structure/seqr/#/
Lewis
SEQR Sequence Search
Home
About
Please input a protein gi, accession number, or sequence (raw/plain text format) to retrieve similar proteins
Search

ALGORITHM

Generate all nmers

GGWAW

PGGPA

GAWGW

PGAPG

PGGPG

AGWGW

PGAPA

GVWGW

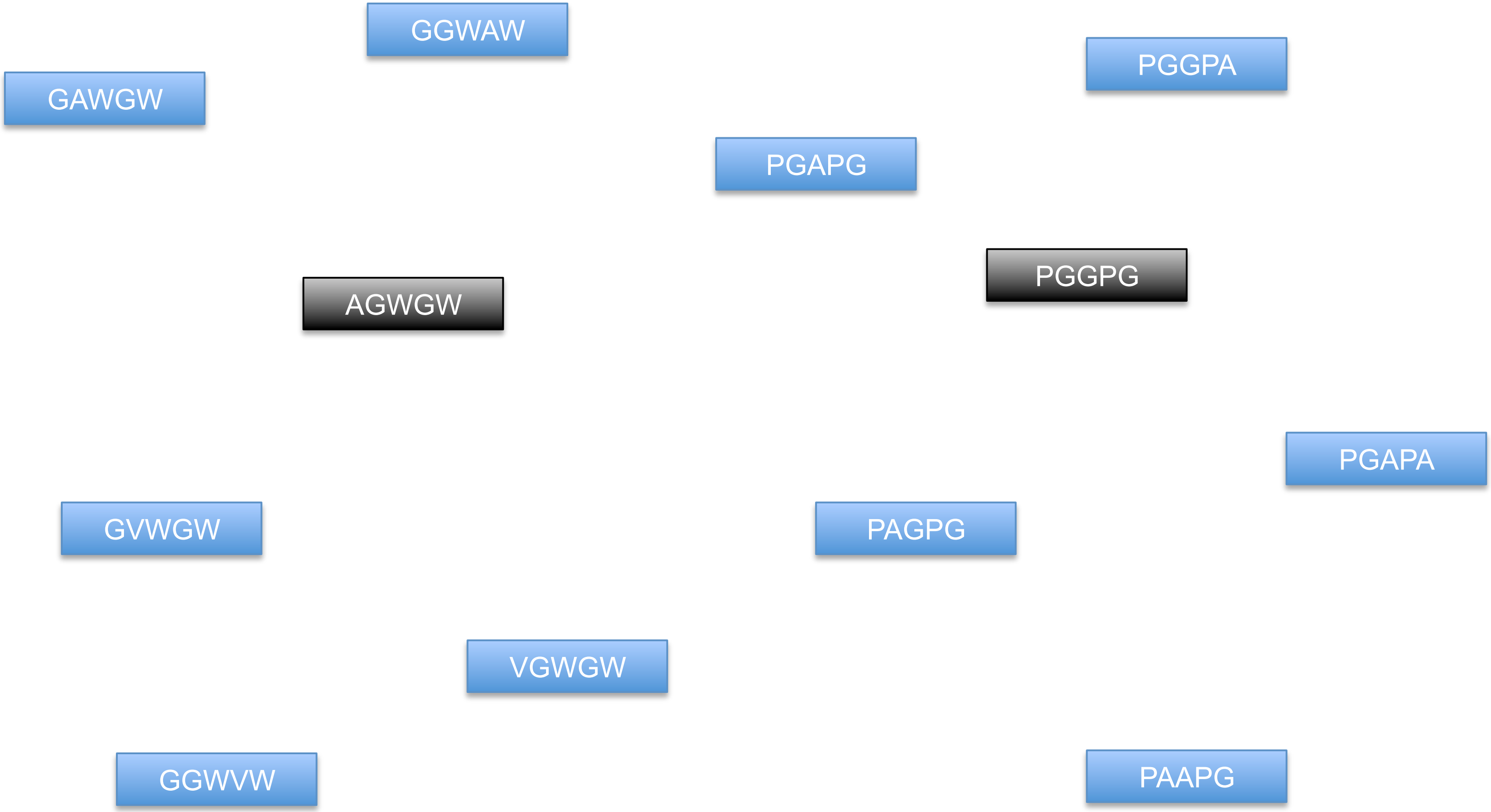
PAGPG

VGWGW

GGWVW

PAAPG

Select cluster seeds



GGWAW

PGGPA

GAWGW

PGAPG

PGGPG

AGWGW

PGAPA

GVWGW

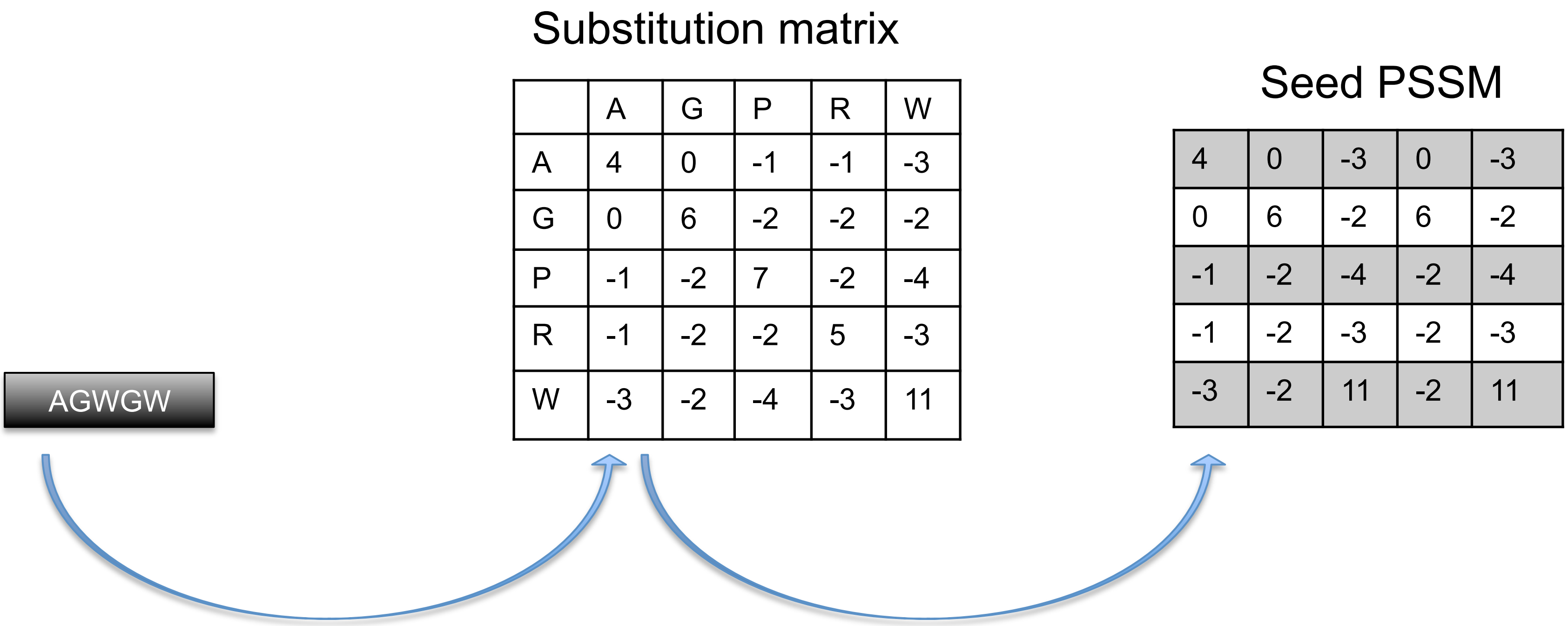
PAGPG

VGWGW

PAAPG

GGWVW

Assign PSSM to seed nmers



AGWGW

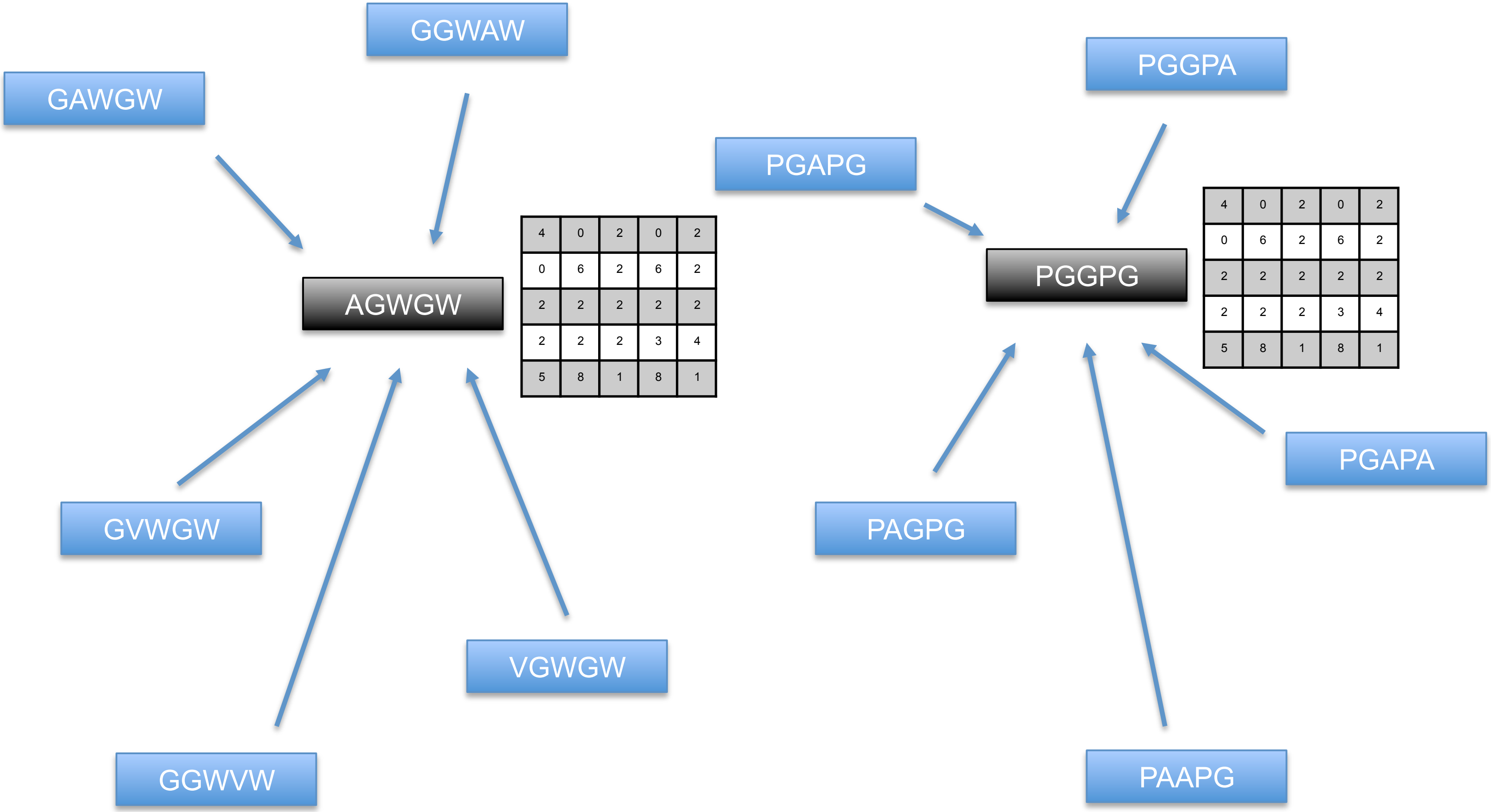
Substitution matrix

	A	G	P	R	W
A	4	0	-1	-1	-3
G	0	6	-2	-2	-2
P	-1	-2	7	-2	-4
R	-1	-2	-2	5	-3
W	-3	-2	-4	-3	11

Seed PSSM

4	0	-3	0	-3
0	6	-2	6	-2
-1	-2	-4	-2	-4
-1	-2	-3	-2	-3
-3	-2	11	-2	11

Cluster all other nmers based on distance to seed PSSM w/ score threshold



4	0	2	0	2
0	6	2	6	2
2	2	2	2	2
2	2	2	3	4
5	8	1	8	1

4	0	2	0	2
0	6	2	6	2
2	2	2	2	2
2	2	2	3	4
5	8	1	8	1

Further steps in clustering

- Seed PSSMs can be updated to average in the PSSMs of merged nmers
- After complete assignment, can reassign nmers based on averaged pssms
- Assign each cluster an ID.

Sequence indexing

- Create an nmer to cluster ID index. Sequences can be indexed in microseconds. This is an optional step.
- Scan each database sequence, position by position. Assign cluster ID to each position either by using index or direct comparison to PSSM.

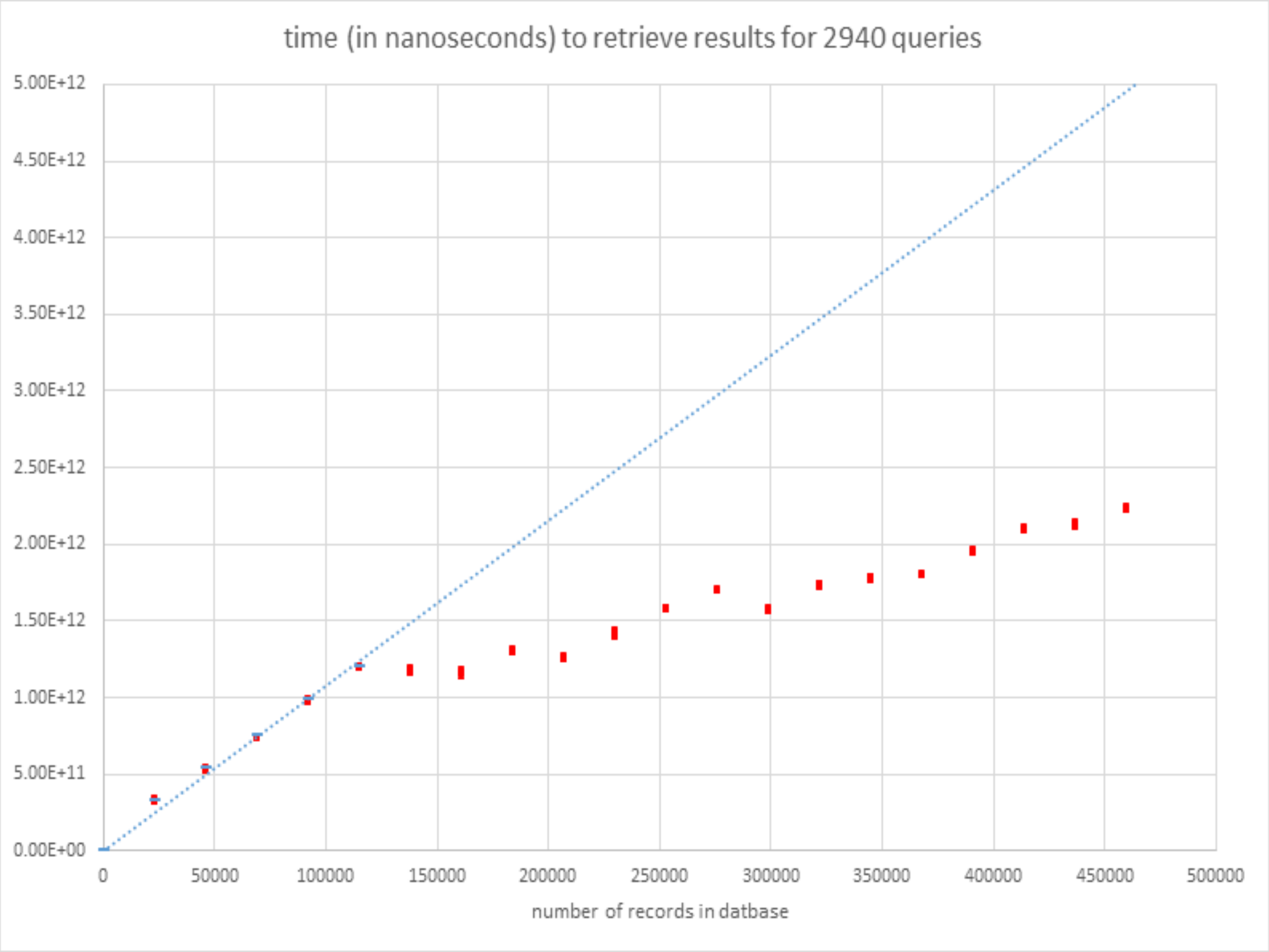
Retrieval

- Create inverted index from cluster ID to sequence ID
- Retrieval procedure:
  - Database is preindexed
  - Query is indexed on the fly
  - For each cluster ID from query, retrieve the sequence IDs and keep track of which sequence IDs have hits

Retrieval, continued…

- After processing all cluster IDs in query, rank hit sequences by Jaccard score: $|A \cap B| / |A| + |B| - |A \cap B|$
- Note that this is a retrieval algorithm, not an alignment algorithm. Alignment can be applied as a post processing step.

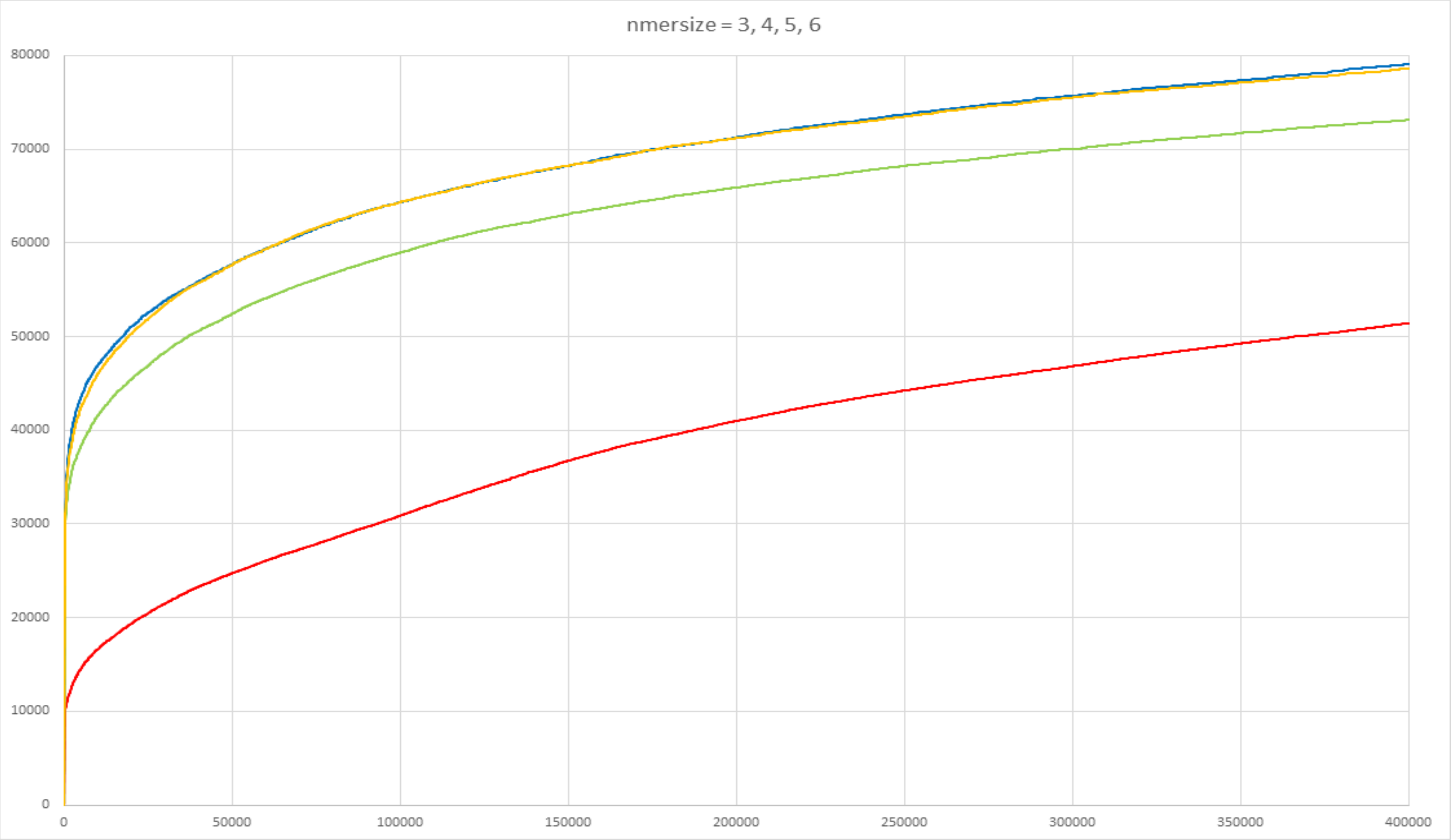
Processing speed



ROC$_n$ analysis

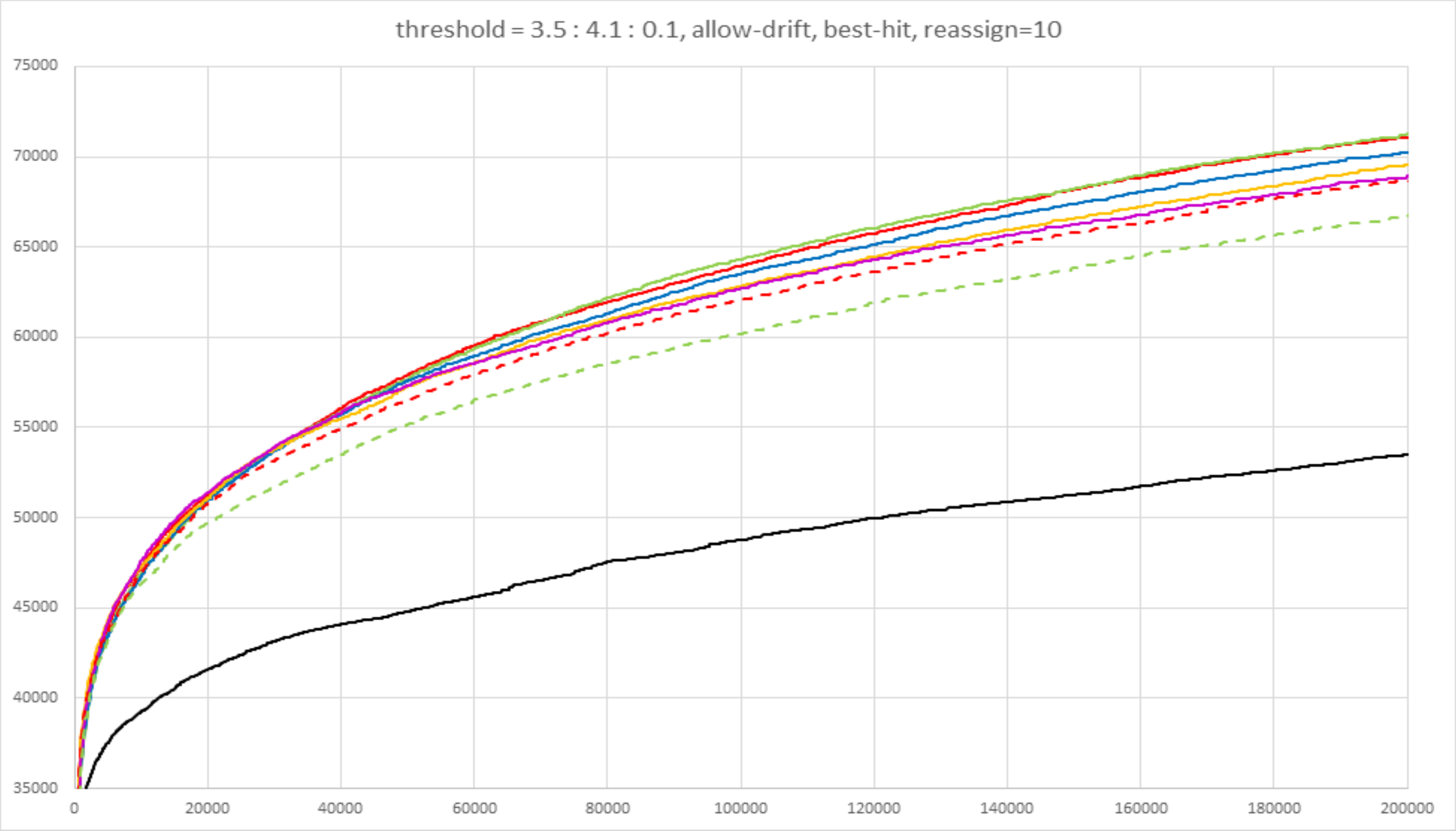
- Compare ~300 query sequences selected from Swiss-Prot, with a bias toward human sequences
- Search against Swiss-Prot (450k sequences)
- Compare to DELTA-BLAST with a threshold cutoff of 1e-4.

Size of nmer



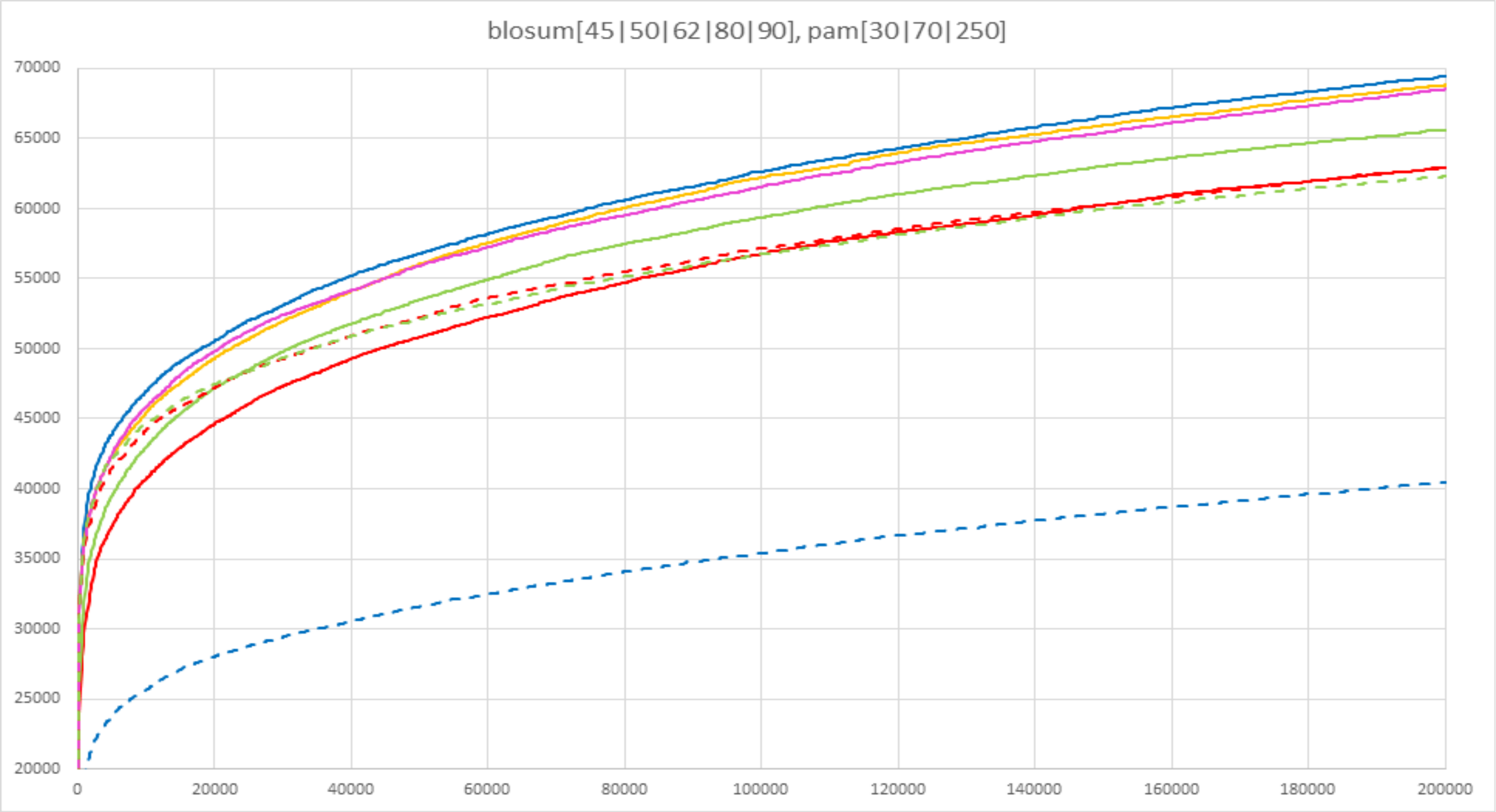
red: n-mer size = 3, threshold = 5.2
green: n-mer size = 4, threshold = 4.6
blue: n-mer size = 5, threshold = 3.6
orange: n-mer size = 6, threshold = 2.8

Vary threshold



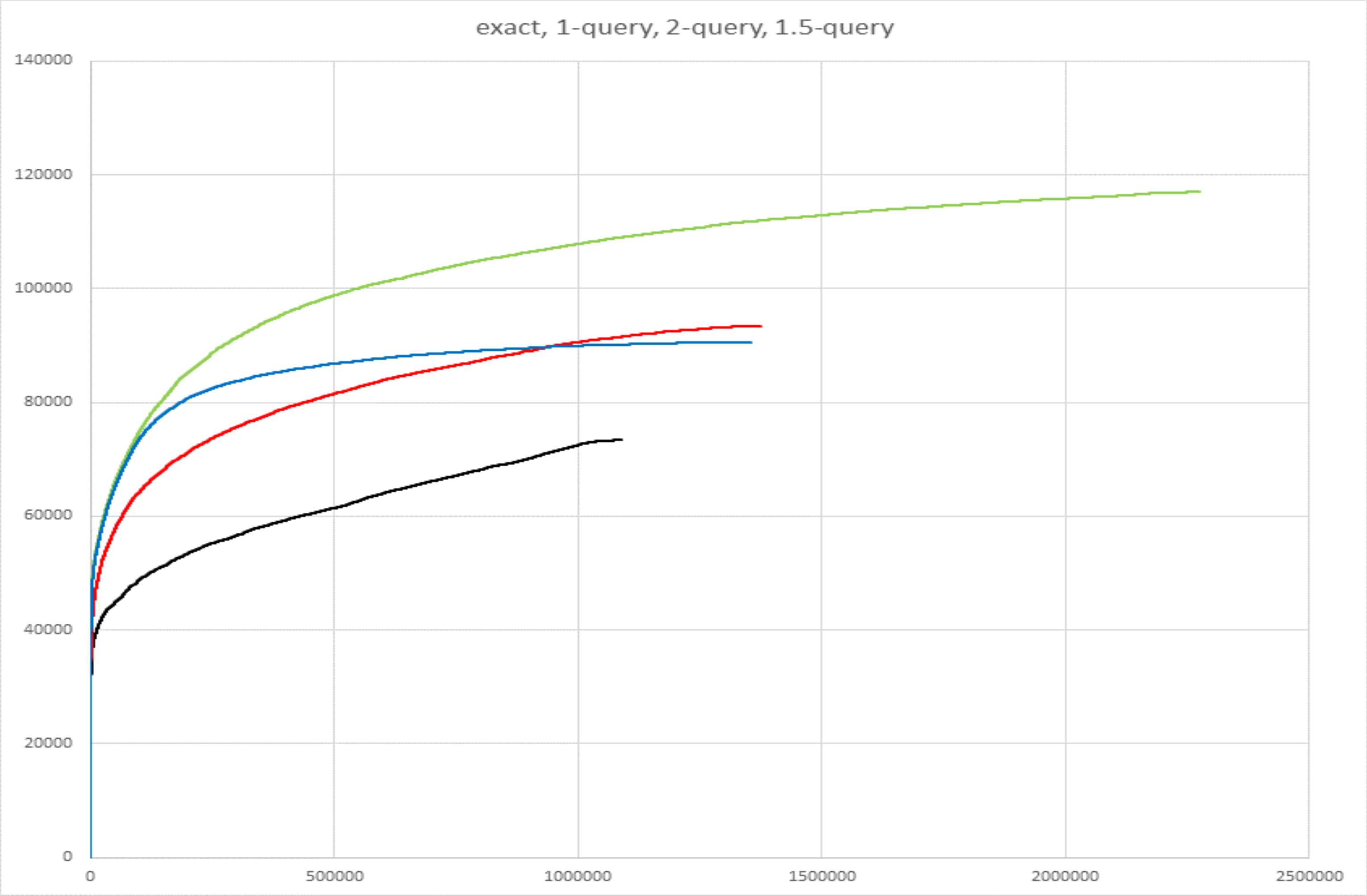
red: 3.5, green: 3.6, blue: 3.7, orange: 3.8, magenta: 3.9
red-dashed: 4.0, green-dashed: 4.1, black: exact

Substitution matrix



red: blosum45, green: blosum50, blue: blosum62, orange: blosum80, magenta: blosum90
red-dashed: pam30, green-dashed: pam70, blue-dashed: pam250

Number of queries



These ROC curves show the improvement of optimized 2-query SEQR (green) vs our best 1-query SEQR (red) and exact-match indices (black)

Cluster sizing

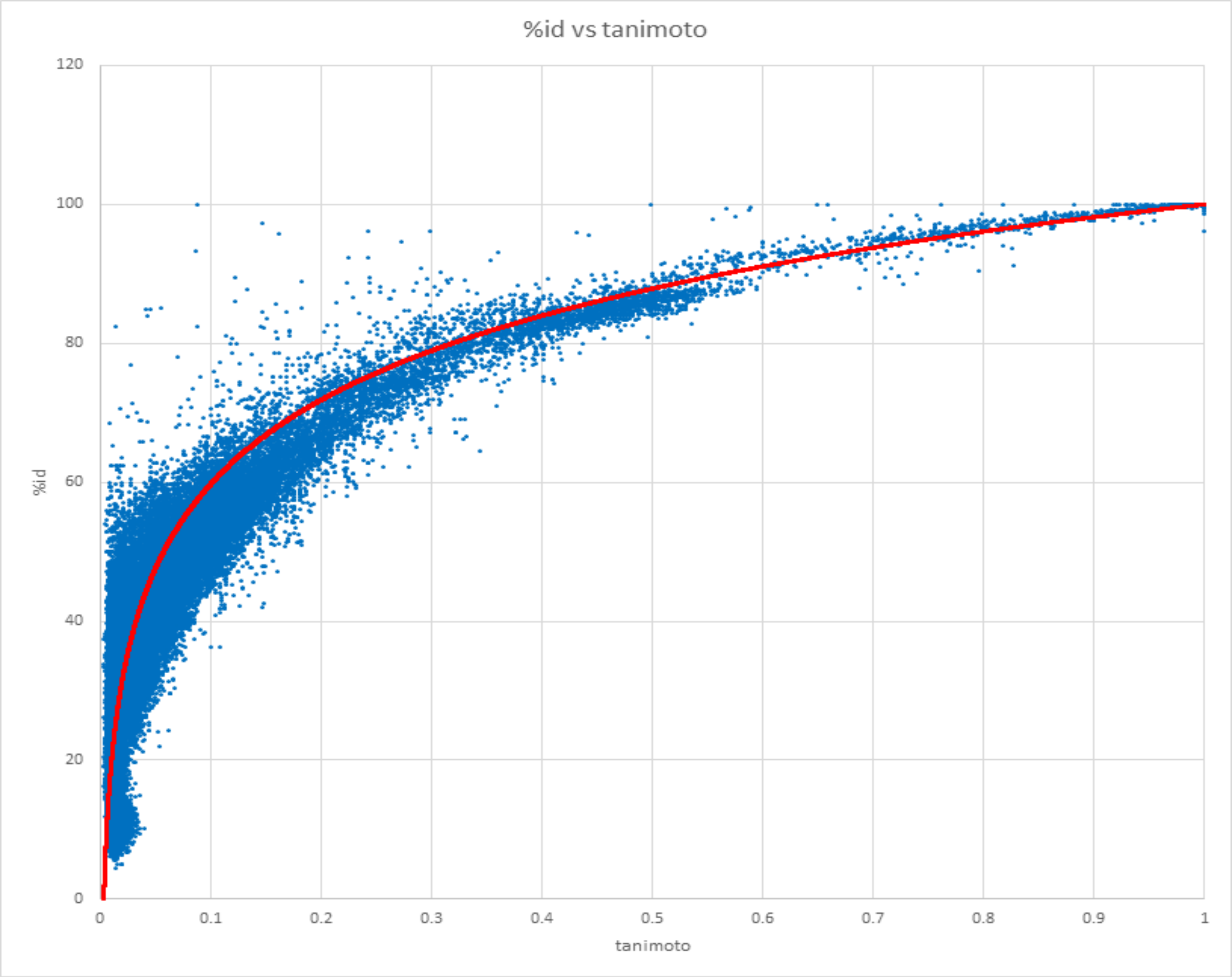
- Best performing indices have ~150k clusters with an average size of ~20

0 -	0 :	5
1 -	1 :	1314
2 -	2 :	4433
3 -	3 :	5785
4 -	4 :	13265
5 -	5 :	6814
6 -	6 :	10643
7 -	7 :	7268
8 -	8 :	14519
9 -	9 :	5922
10 -	19 :	42642
20 -	29 :	14647
30 -	39 :	7724
40 -	49 :	4336
50 -	59 :	2407
60 -	69 :	2149
70 -	79 :	1274
80 -	89 :	984
90 -	99 :	815
100 -	199 :	3131
200 -	299 :	924
300 -	399 :	367
400 -	499 :	72
500 -	599 :	47
600 -	699 :	25
700 -	799 :	16
800 -	899 :	11
900 -	999 :	11
1000 -	1999 :	11

Sample cluster, seed FFTLT

FFTLT
FFTLI
FFTVT
FFVLT
FFKLT
YFTMT
FFTYT
FFRLT
FFSLT
FFTMT
FFTLQ
FFTLM
FYTLT
FFTLS
FFTLN
FFTLP
YFTLT
FFALT
FFTAT
FFTFT
FYTMT
FYTIT
FFTCT
FFTLV
FFTLL
FFELT
FFTTT
FFTIT

Compare BLAST hits to SEQR



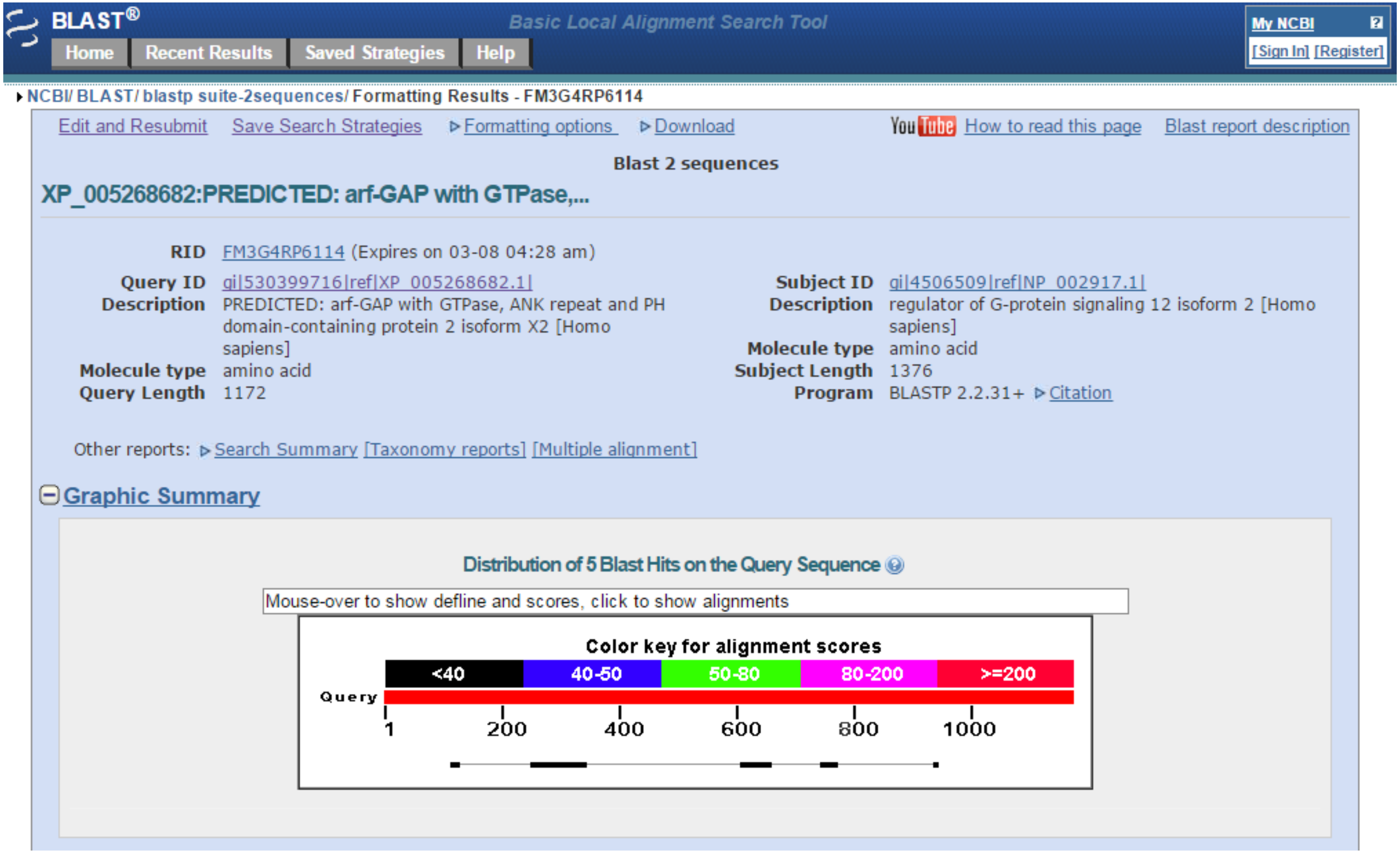
Example hit

```
Query  185  LQRHRIMHRGDGPYKCKFCGKALMFLSLYLIHKRTHTGEK-------------------P  225
            L  +++  +GD     K C K     + +    +  TGEK                   P
Sbjct  195  LPNNKLSDKGDKNQTSKKCEKVCRHSASHTKEDKIQTGEKRKSHCRTPSKPEKAPGSGKP  254

Query  226  YQCKQCGKAFSHSSSLRIHERTHTGEKPYKCNECGKAFHSSTCLHAHKRTHTGEKPYECK  285
            Y+C  CGK  SH   L  H+RTHTGEKPY+CNECG AF   + L  H+RTHTGEKPYEC+
Sbjct  255  YECNHCGKVLSHKQGLLDHQRTHTGEKPYECNECGIAFSQKSHLVVHQRTHTGEKPYECE  314

Query  286  QCGKAFSSSHSFQIHERTHTGEKPYECKECGKAFKCPSSVRRHERTHSRKKPYECKHCGK  345
            QCGKA    H+   H R HTGEKPY+C ECGK F+  S++ +H R+H+ +KPYECK CGK
Sbjct  315  QCGKAHGHKHALTDHLRIHTGEKPYKCNECGKTFRHSSNLMQHLRSHTGEKPYECKECGK  374

Query  346  VLSYLTSFQNHLGMHTGEISHKCKICGKAFYSPSSLQTHEKTHTGEKPYKCNQCGKAFNS  405
               Y +S   H+  HTGEI ++C  CGKAF   SSL  H + HTGEKP++CN+CGK F+
Sbjct  375  SFRYNSSLTEHVRTHTGEIPYECNECGKAFKYGSSLTKHMRIHTGEKPFECNECGKTFSK  434

Query  406  SSSFRYHERTHTGEKPYECKQCGKAFRSASLLQTHGRTHTGEKPYACKECGKPFSNFSFF  465
             S    H+RTHT EKPY+C +CGKAF  +S L  H RTHTG+ P+ C +CGK F
Sbjct  435  KSHLVIHQRTHTKEKPYKCDECGKAFGHSSSLTYHMRTHTGDCPFECNQCGKAFKQIEGL  494

Query  466  QIHERMHREEKPYECKGYGKTFSLPSLFHRHERTHTGGKTYECKQCGRSFNCSSSFRYHG  525
              H+R+H  EKPYEC   GK FS  S    H+RTHTG K +EC +CG++FN  S    H
Sbjct  495  TQHQRVHTGEKPYECVECGKAFSQKSHLIVHQRTHTGEKPFECYECGKAFNAKSQLVIHQ  554

Query  526  RTHTGEKPYECKQCGKAFRSASQLQIHGRTHTGEKPYE  563
            R+HTGEKPYEC +CGKAF+  + L  H + H+ E+  E
Sbjct  555  RSHTGEKPYECIECGKAFKQNASLTKHMKIHSEEQSEE  592
```

- BLASTP percent identity: 37%
- SEQR tanimoto: 0.2

- 213 SEQR indexing terms
- For identical n-mers, 34 indexing terms

Typical hit found by SEQR but not by blastp



NCBI/ BLAST/ blastp suite-2sequences/ Formatting Results - FM3G4RP6114

Blast 2 sequences

XP_005268682:PREDICTED: arf-GAP with GTPase,...

RID FM3G4RP6114 (Expires on 03-08 04:28 am)

Query ID gi|530399716|ref|XP_005268682.1|
Description PREDICTED: arf-GAP with GTPase, ANK repeat and PH domain-containing protein 2 isoform X2 [Homo sapiens]
Molecule type amino acid
Query Length 1172

Subject ID gi|4506509|ref|NP_002917.1|
Description regulator of G-protein signaling 12 isoform 2 [Homo sapiens]
Molecule type amino acid
Subject Length 1376
Program BLASTP 2.2.31+ Citation

Other reports: Search Summary [Taxonomy reports] [Multiple alignment]

Graphic Summary

Distribution of 5 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments

IMPLEMENTATION

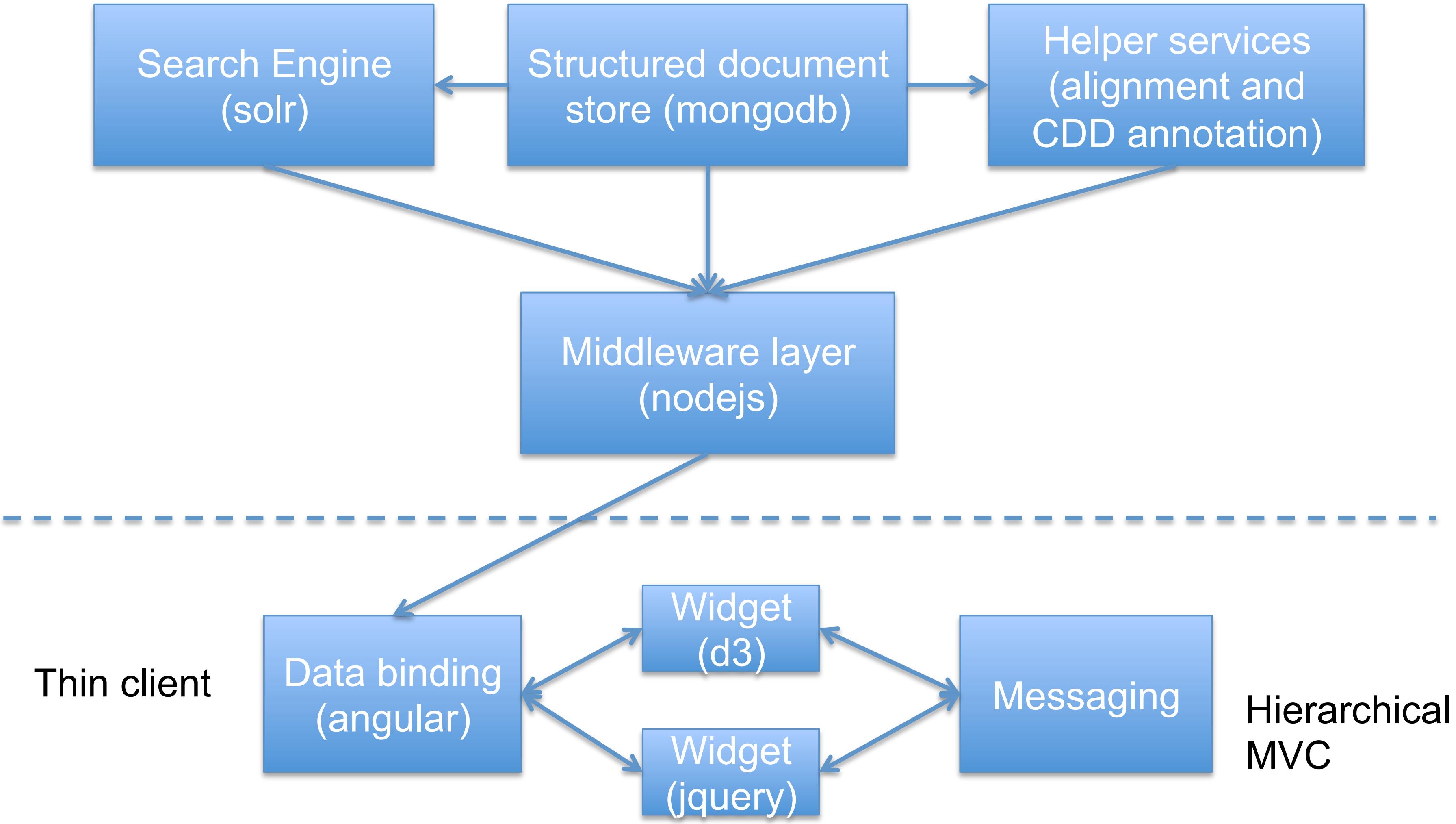
Implementation

- In the past few years, a quiet revolution in open source projects for information retrieval:
  - Search engines: solr, elasticsearch, …
  - Structured data stores using JSON: mongodb, couchbase, …
  - Middleware architectures with extensive libraries: nodejs, django, …
  - Javascript: jquery, d3, angular
  - Even c++ has benefited: boost libraries, cereal, …

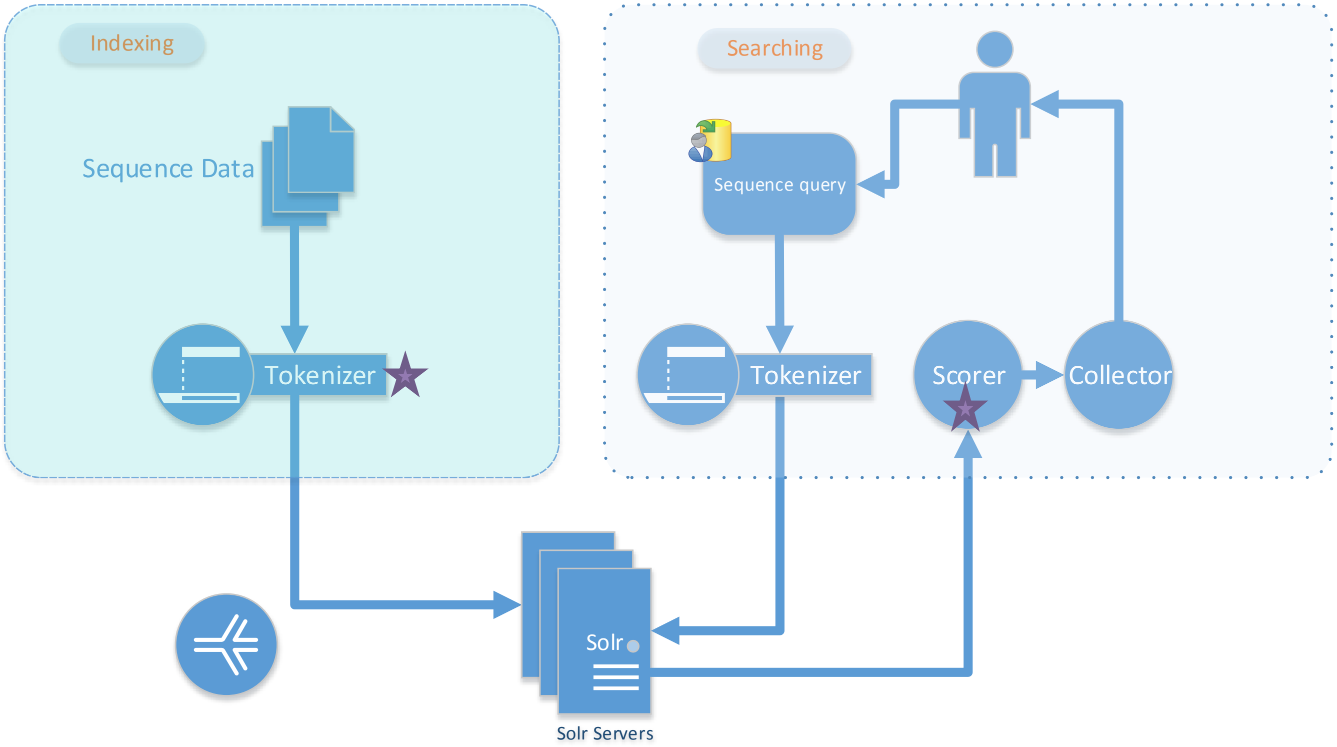
Implementation

- Web 2.0. Clean implementation of thin clients. The backend is SaaS.
- Can we use these libraries and techniques to speed up the development process (or make it possible for small groups to do big things)?

Architecture


Search Engine (solr)
Structured document store (mongodb)
Helper services (alignment and CDD annotation)
Middleware layer (nodejs)
Thin client
Data binding (angular)
Widget (d3)
Widget (jquery)
Messaging
Hierarchical MVC

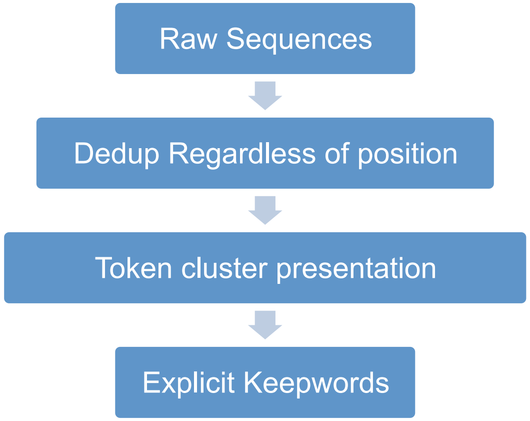
Workflow


Indexing
Sequence Data
Tokenizer
Searching
Sequence query
Tokenizer
Scorer
Collector
Solr
Solr Servers

Tokenizer

Implementation

```
<fieldType name="sequence" class="solr.TextField" >
    <analyzer>
        <tokenizer class="gov.nih.nlm.ncbi.seqr.tokenizer.RawSequence2TokenizerFactory" nmer="5"/>
        <filter class="gov.nih.nlm.ncbi.seqr.tokenizer.RemoveDuplicatesTokenFilterIgnorePositionFactory"/>
        <filter class="solr.SynonymFilterFactory" synonyms="seqr.txt" ignoreCase="true" expand="false"/>
        <filter class="solr.KeepWordFilterFactory" words="seqr.keeplist" ignoreCase="true"/>
    </analyzer>
</fieldType>
```



https://github.com/NCBI-Hackathons/seqr-tokenizer

Scorer

Tanimoto/Jaccard Index

```
<fieldType name="sequence" class="solr.TextField" >
    <similarity class="org.apache.lucene.analysis.tanimoto.OverlapSimilarity"/>
    <analyzer>
        <tokenizer class="gov.nih.nlm.ncbi.seqr.tokenizer.RawSequence2TokenizerFactory" nmer="5"/>
        <filter class="gov.nih.nlm.ncbi.seqr.tokenizer.RemoveDuplicatesTokenFilterIgnorePositionFactory"/>
        <filter class="solr.SynonymFilterFactory" synonyms="seqr.txt" ignoreCase="true" expand="false"/>
        <filter class="solr.KeepWordFilterFactory" words="seqr.keeplist" ignoreCase="true"/>
    </analyzer>
</fieldType>
```

```
q={!tanimoto bf=seqLen v=$qq}
qq={!edismax mm=50% qf=sequence}(bit1 bit2)
```

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

https://github.com/NCBI-Hackathons/solr-tanimoto

Hardware

- Search running on one 2 CPU Dell C6220, 256 GB RAM, 1 TB SSD – low end commodity server
- ¼ of a 2U chassis.
- >200M sequences indexed. ~100 GB index size (< 1k/sequence)
- 15 cores used
- Typical 60 GB heap usage

Loading time

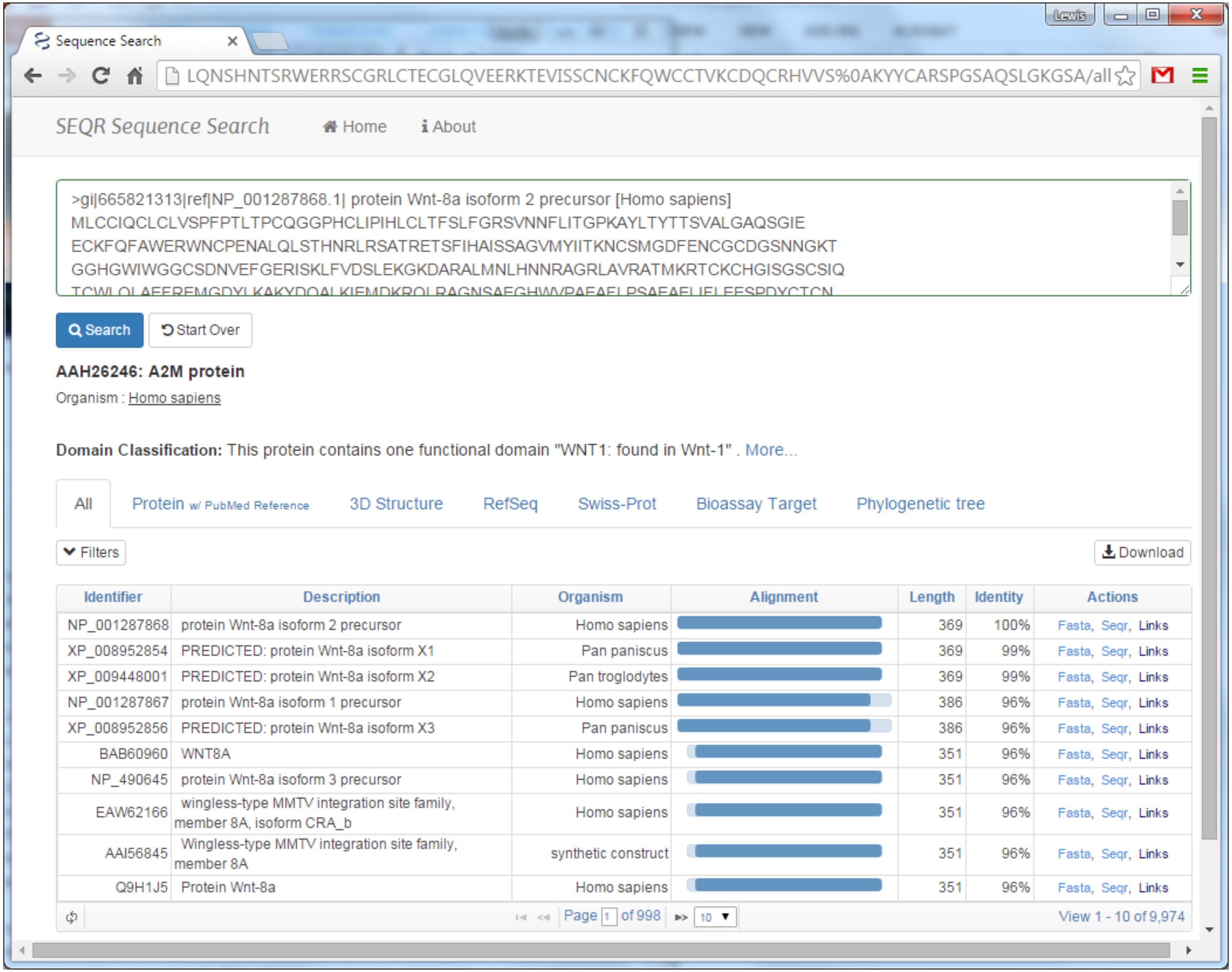
- Batch update of 236M records
  - Load MongoDB: 10.5 hours (1 thread). Limited by use of one SQL-Server.
  - Dump MongoDB: 5 hours (20 threads). Limited by number of MongoDB servers being queried (1).
  - Load Solr: 2.5 hours (16 threads).
  - Total time is 18 hours.

Acknowledgements

Lianyi Han
Shennan Lu
Jane He
Steve Bryant
Stephen Altschul
David Hurwitz
Bo Yu
Aron Marchler-Bauer
David Lipman

Future work

- Further analysis of algorithm
  - Clustering improvements
- Use of cluster pssms derived from CDD
  - Extension to domain annotation
- Extension to DNA sequences
- Alignment generation


Sequence Search
LQNSHNTSRWERRSCGRLCTECGLQVEERKTEVISSCNCKFQWCCTVKCDQCRHVVS%0AKYYCARSPGSAQSLGKGSA/all
Lewis
SEQR Sequence Search
Home
About
Search
Start Over

```
>gi|665821313|ref|NP_001287868.1| protein Wnt-8a isoform 2 precursor [Homo sapiens]
MLCCIQCLCLVSPFPTLTPCQGGPHCLIPIHLCLTFSLFGRSVNNFLITGPKAYLTYTTSVALGAQSGIE
ECKFQFAWERWNCPENALQLSTHNRLRSATRETSFIHAISSAGVMYIITKNCSMGDFENCGCDGSNNGKT
GGHGWIWGGCSDNVEFGERISKLFVDSLEKGKDARALMNLHNNRAGRLAVRATMKRTCKCHGISGSCSIQ
TCWLQLAEFREMGDYLKAKYDQALKIEMDKRQLRAGNSAEGHWVPAEAFLPSAEAELIFLEESPDYCTCN
```

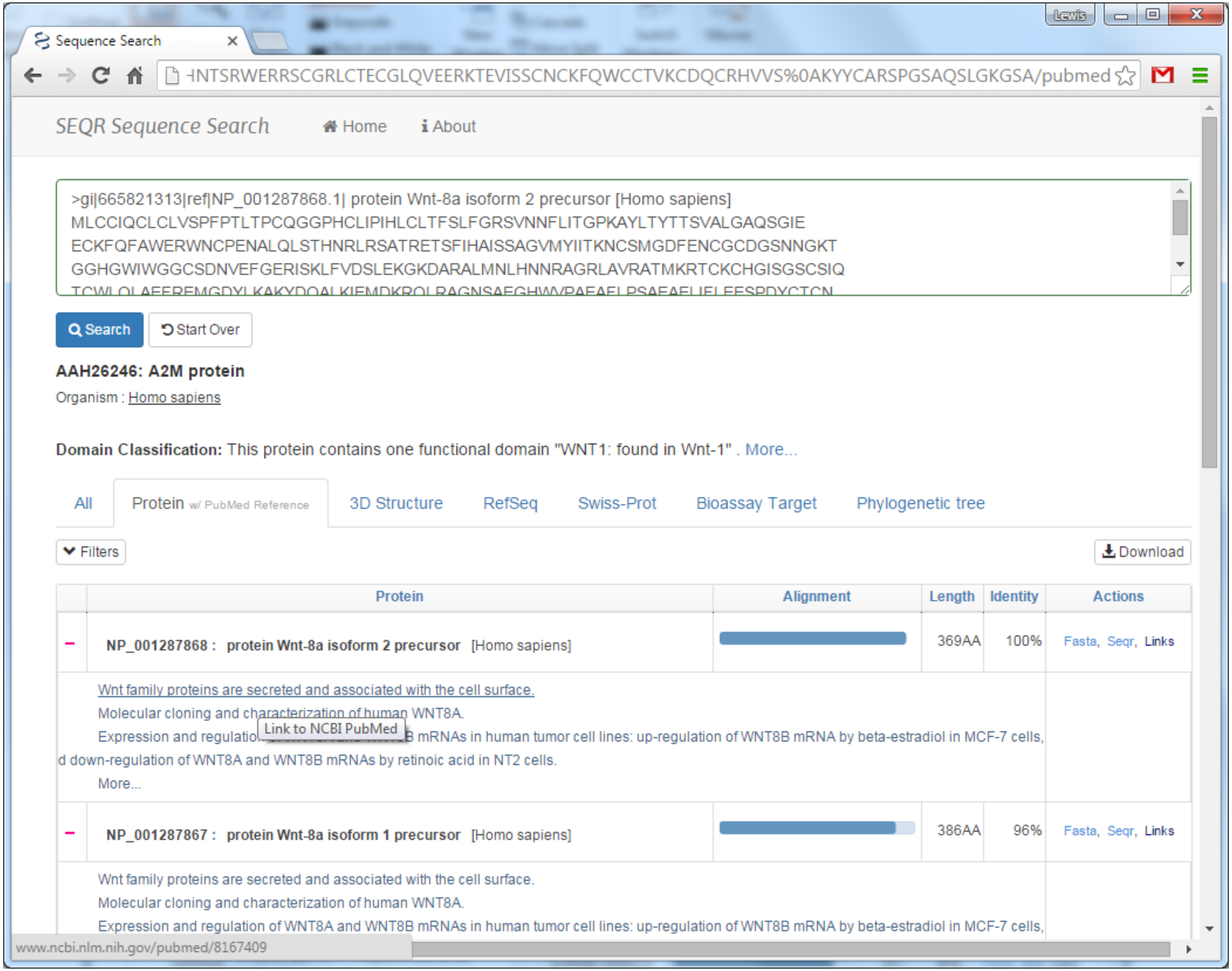
AAH26246: A2M protein

Organism : Homo sapiens

Domain Classification: This protein contains one functional domain "WNT1: found in Wnt-1" . More...

Identifier	Description	Organism	Alignment	Length	Identity	Actions
NP_001287868	protein Wnt-8a isoform 2 precursor	Homo sapiens		369	100%	Fasta, Seqr, Links
XP_008952854	PREDICTED: protein Wnt-8a isoform X1	Pan paniscus		369	99%	Fasta, Seqr, Links
XP_009448001	PREDICTED: protein Wnt-8a isoform X2	Pan troglodytes		369	99%	Fasta, Seqr, Links
NP_001287867	protein Wnt-8a isoform 1 precursor	Homo sapiens		386	96%	Fasta, Seqr, Links
XP_008952856	PREDICTED: protein Wnt-8a isoform X3	Pan paniscus		386	96%	Fasta, Seqr, Links
BAB60960	WNT8A	Homo sapiens		351	96%	Fasta, Seqr, Links
NP_490645	protein Wnt-8a isoform 3 precursor	Homo sapiens		351	96%	Fasta, Seqr, Links
EAW62166	wingless-type MMTV integration site family, member 8A, isoform CRA_b	Homo sapiens		351	96%	Fasta, Seqr, Links
AAI56845	Wingless-type MMTV integration site family, member 8A	synthetic construct		351	96%	Fasta, Seqr, Links
Q9H1J5	Protein Wnt-8a	Homo sapiens		351	96%	Fasta, Seqr, Links

Page 1 of 998 10 ▼ View 1 - 10 of 9,974



-INTSRWERRSCGRLCTECGLQVEERKTEVISSCNCKFQWCCTVKCDQCRHVVS%0AKYYCARSPGSAQSLGKGSA/pubmed

```
>gi|665821313|ref|NP_001287868.1| protein Wnt-8a isoform 2 precursor [Homo sapiens]
MLCCIQCLCLVSPFPTLTPCQGGPHCLIPIHLCLTFSLFGRSVNNFLITGPKAYLTYTTSVALGAQSGIE
ECKFQFAWERWNCPENALQLSTHNRLRSATRETSFIHAISSAGVMYIITKNCSMGDFENCGCDGSNNGKT
GGHGWIWGGCSDNVEFGERISKLFVDSLEKGKDARALMNLHNNRAGRLAVRATMKRTCKCHGISGSCSIQ
TCWLQLAEFREMGDYLKAKYDQALKIEMDKRQLRAGNSAFGHWVPAEALPSAEAELIFLEESPDYCTCN
```

AAH26246: A2M protein

Organism : Homo sapiens

Domain Classification: This protein contains one functional domain "WNT1: found in Wnt-1" . More...

All | Protein w/ PubMed Reference | 3D Structure | RefSeq | Swiss-Prot | Bioassay Target | Phylogenetic tree

	Protein	Alignment	Length	Identity	Actions
−	**NP_001287868 : protein Wnt-8a isoform 2 precursor** [Homo sapiens]		369AA	100%	Fasta, Seqr, Links

Wnt family proteins are secreted and associated with the cell surface.

Molecular cloning and characterization of human WNT8A.

Expression and regulatio Link to NCBI PubMed B mRNAs in human tumor cell lines: up-regulation of WNT8B mRNA by beta-estradiol in MCF-7 cells, d down-regulation of WNT8A and WNT8B mRNAs by retinoic acid in NT2 cells.

More...

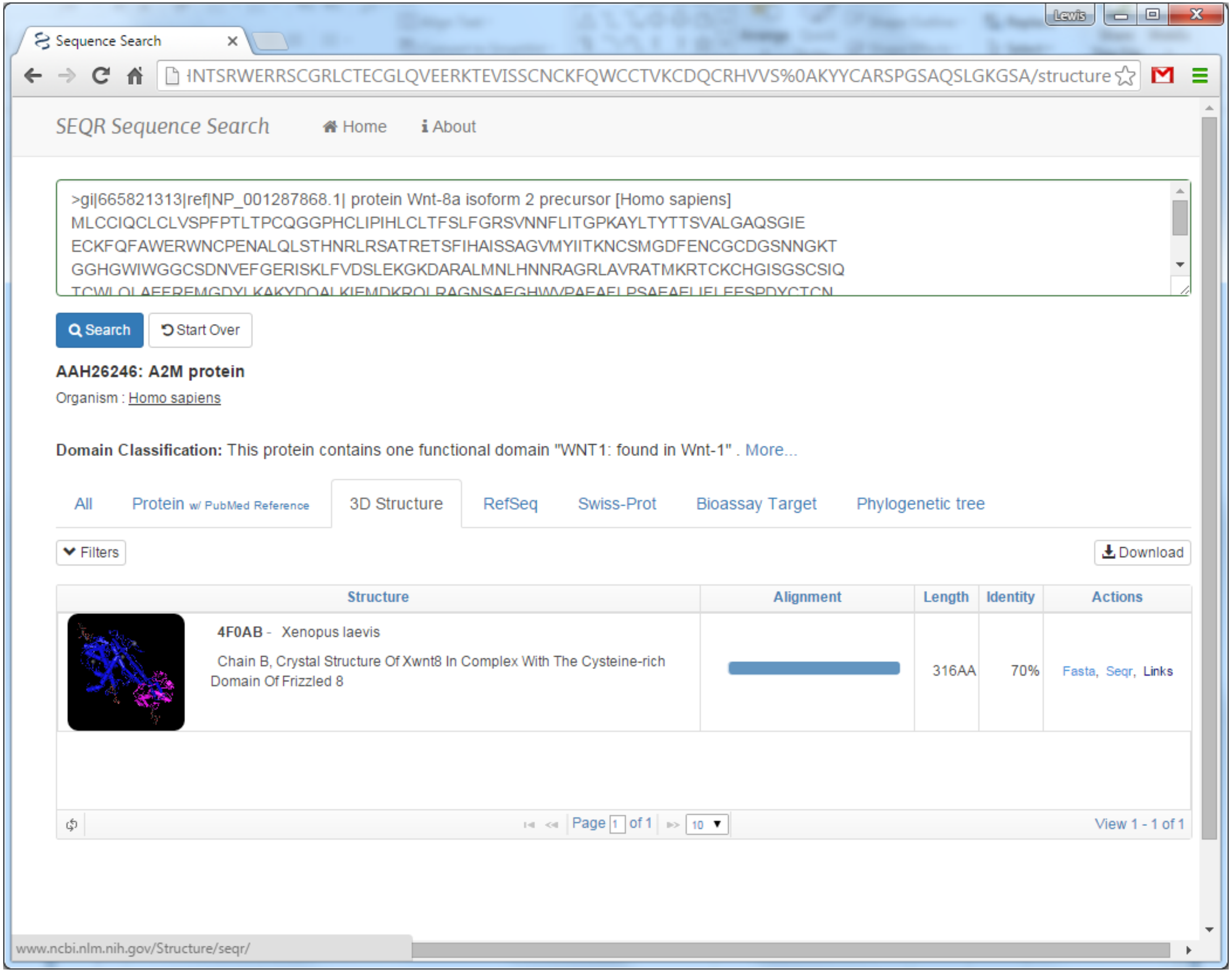
	Protein	Alignment	Length	Identity	Actions
−	**NP_001287867 : protein Wnt-8a isoform 1 precursor** [Homo sapiens]		386AA	96%	Fasta, Seqr, Links

Wnt family proteins are secreted and associated with the cell surface.

Molecular cloning and characterization of human WNT8A.

Expression and regulation of WNT8A and WNT8B mRNAs in human tumor cell lines: up-regulation of WNT8B mRNA by beta-estradiol in MCF-7 cells,

www.ncbi.nlm.nih.gov/pubmed/8167409



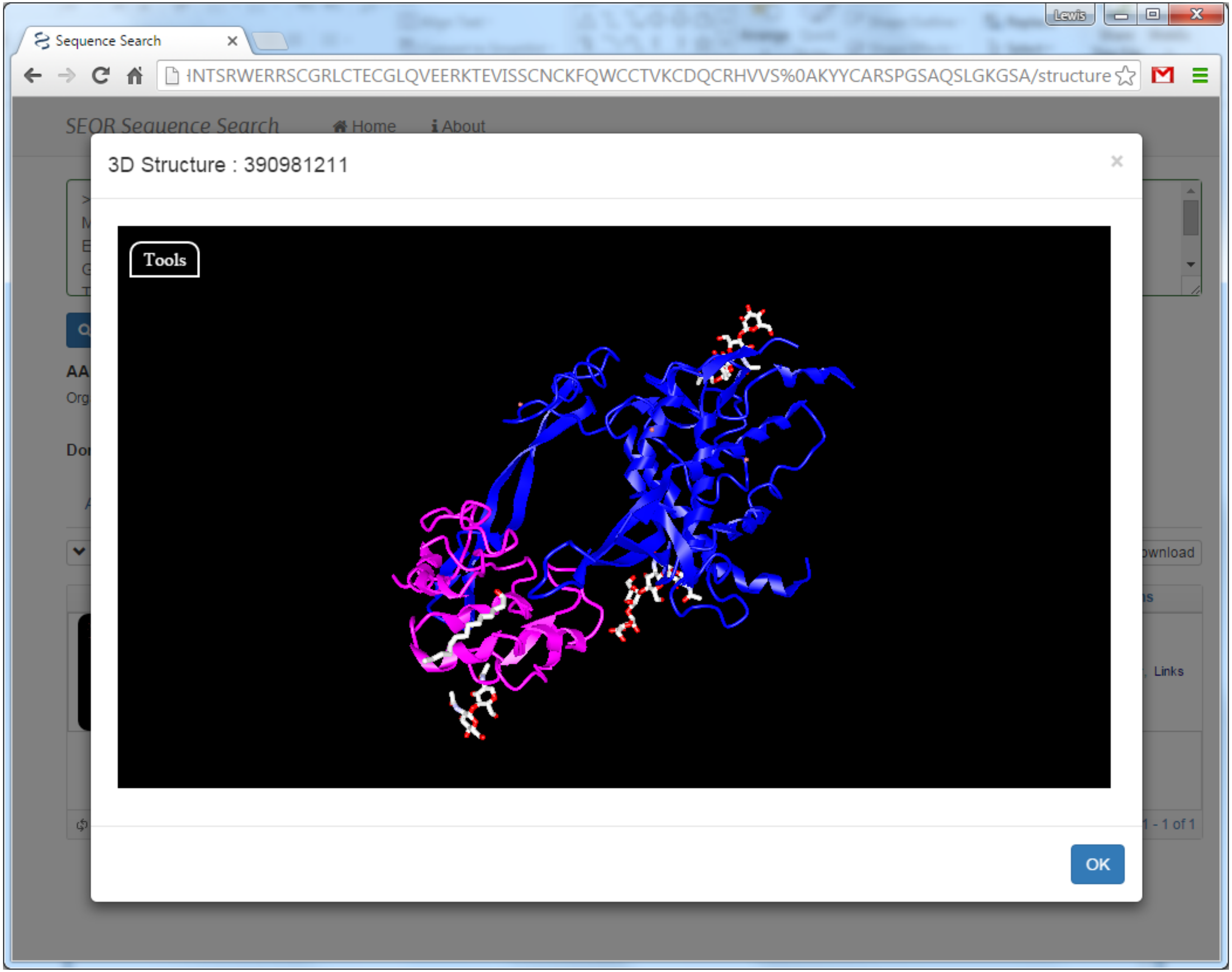
```
>gi|665821313|ref|NP_001287868.1| protein Wnt-8a isoform 2 precursor [Homo sapiens]
MLCCIQCLCLVSPFPTLTPCQGGPHCLIPIHLCLTFSLFGRSVNNFLITGPKAYLTYTTSVALGAQSGIE
ECKFQFAWERWNCPENALQLSTHNRLRSATRETSFIHAISSAGVMYIITKNCSMGDFENCGCDGSNNGKT
GGHGWIWGGCSDNVEFGERISKLFVDSLEKGKDARALMNLHNNRAGRLAVRATMKRTCKCHGISGSCSIQ
TCWLQLAEFREMGDYLKAKYDQALKIEMDKRQLRAGNSAEGHWVPAEAFLPSAEAELIFLEESPDYCTCN
```

AAH26246: A2M protein

Organism : Homo sapiens

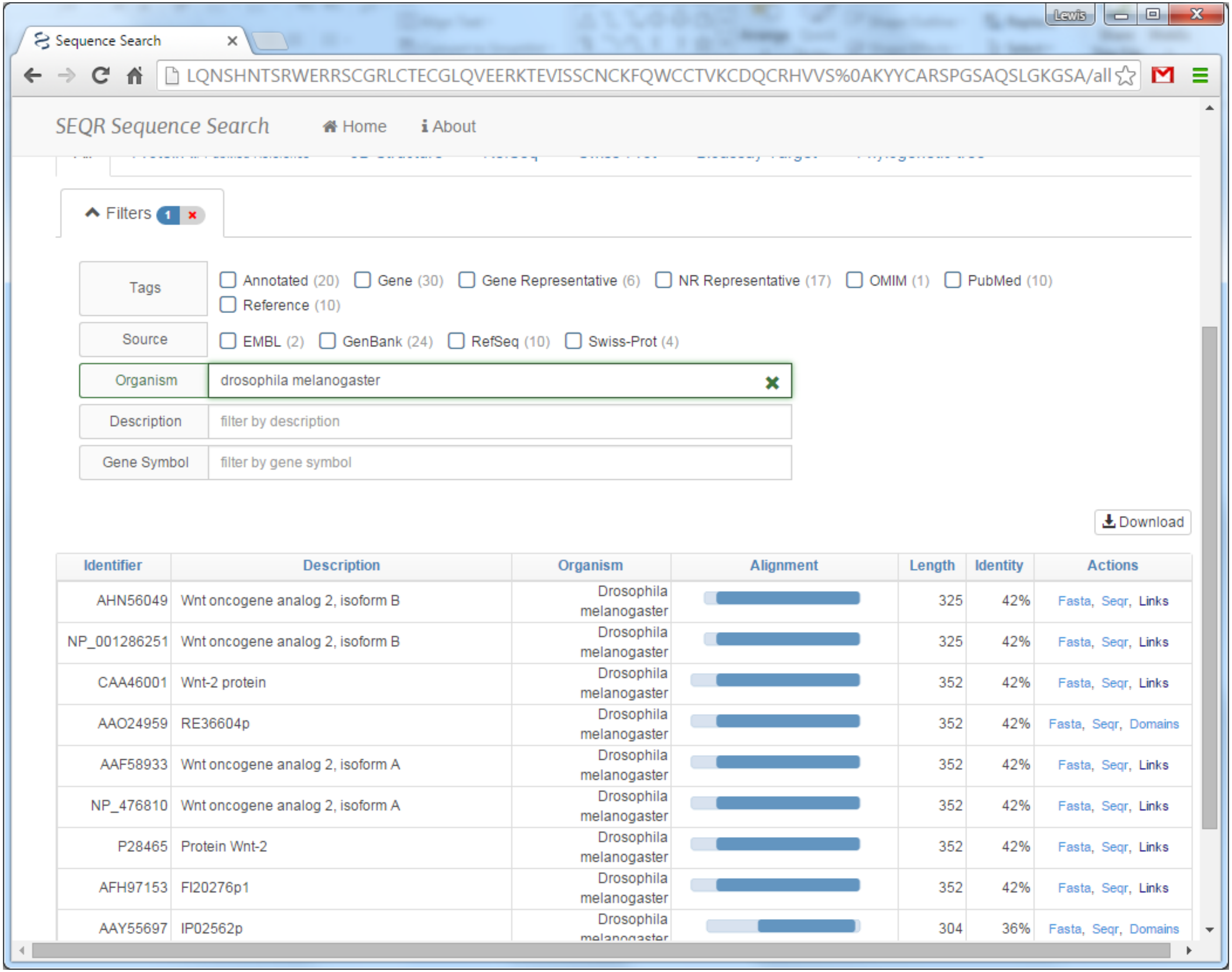
Domain Classification: This protein contains one functional domain "WNT1: found in Wnt-1" . More...

Structure	Alignment	Length	Identity	Actions
4F0AB - Xenopus laevis Chain B, Crystal Structure Of Xwnt8 In Complex With The Cysteine-rich Domain Of Frizzled 8		316AA	70%	Fasta, Seqr, Links



ıNTSRWERRSCGRLCTECGLQVEERKTEVISSCNCKFQWCCTVKCDQCRHVVS%0AKYYCARSPGSAQSLGKGSA/structure

3D Structure : 390981211


Sequence Search
LQNSHNTSRWERRSCGRLCTECGLQVEERKTEVISSCNCKFQWCCTVKCDQCRHVVS%0AKYYCARSPGSAQSLGKGSA/all
SEQR Sequence Search
Home
About
Filters 1 ×
Download

Tags: □ Annotated (20) □ Gene (30) □ Gene Representative (6) □ NR Representative (17) □ OMIM (1) □ PubMed (10) □ Reference (10)

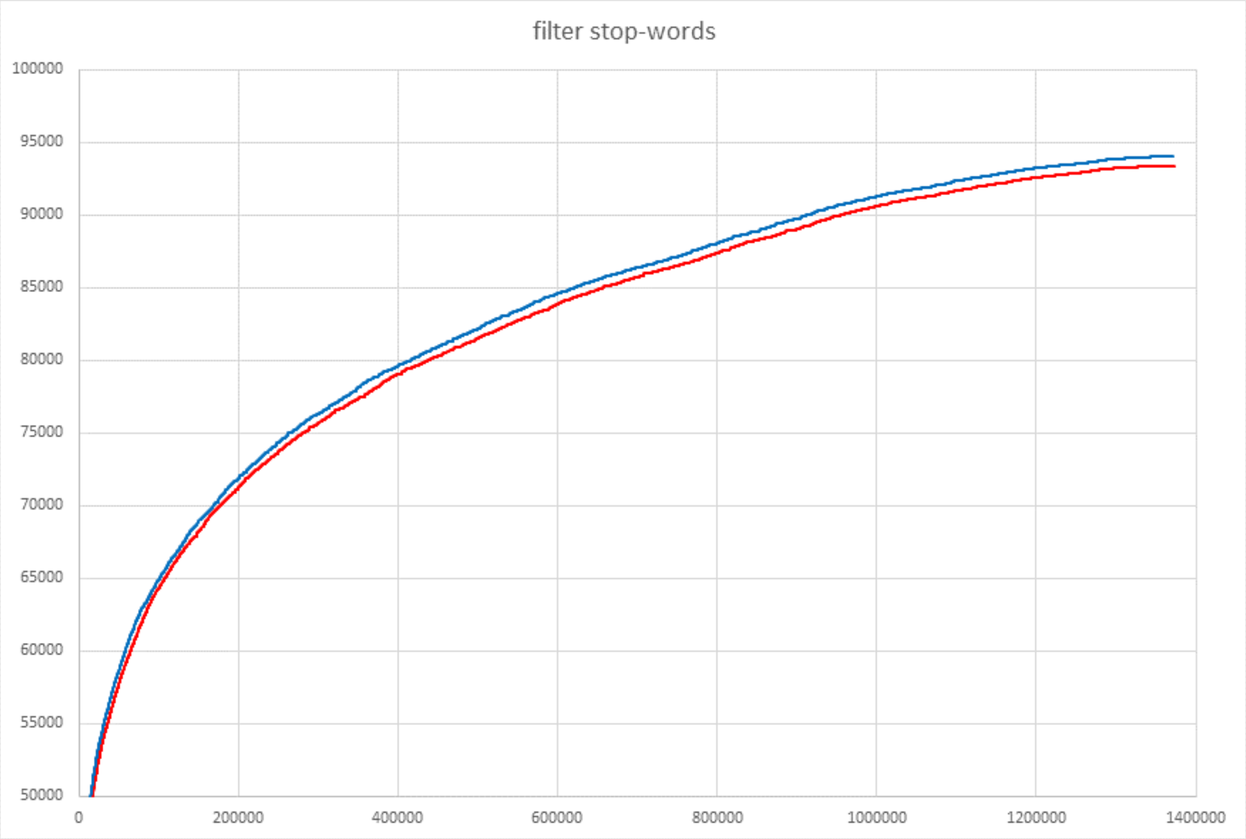
Source: □ EMBL (2) □ GenBank (24) □ RefSeq (10) □ Swiss-Prot (4)

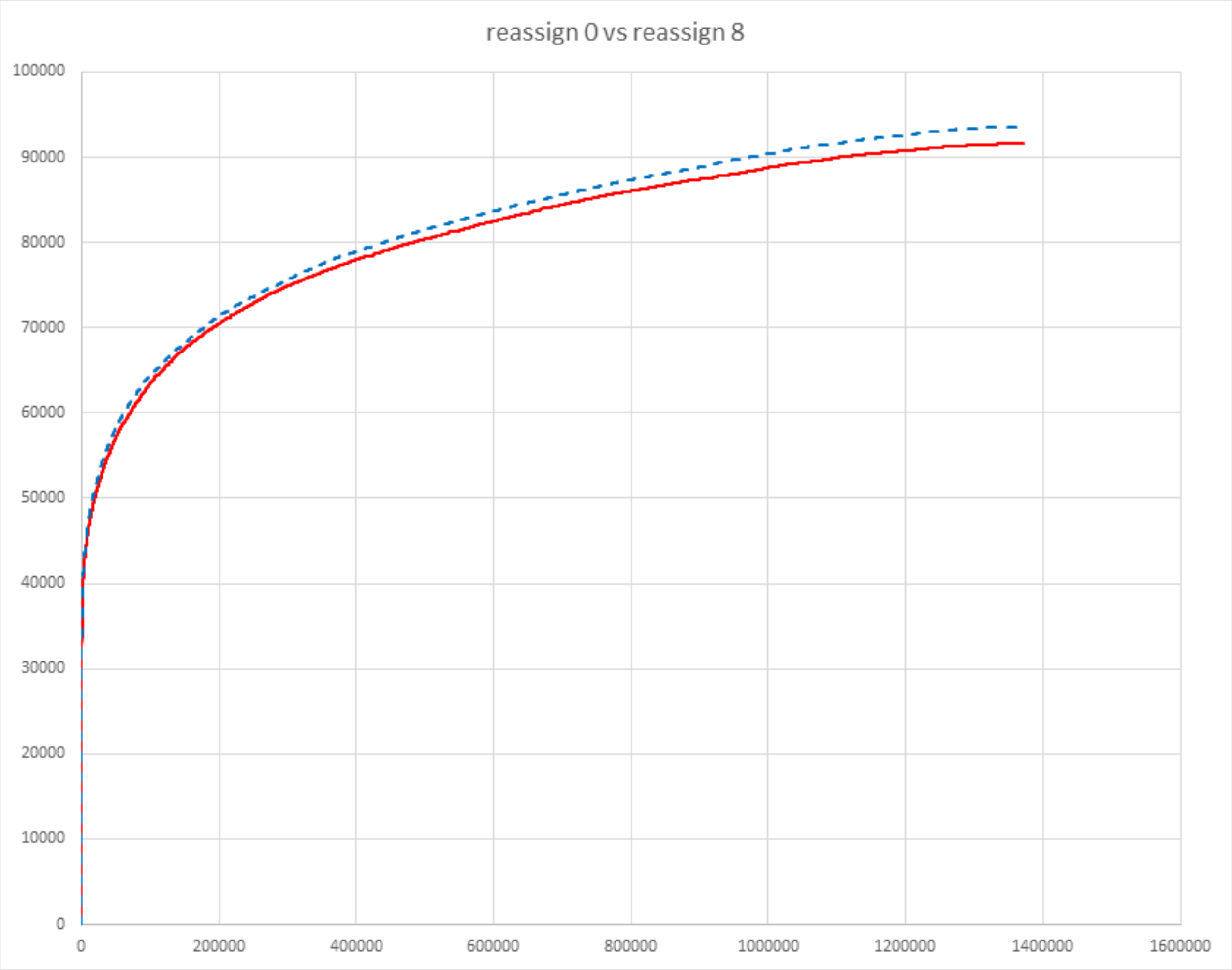
Organism: drosophila melanogaster

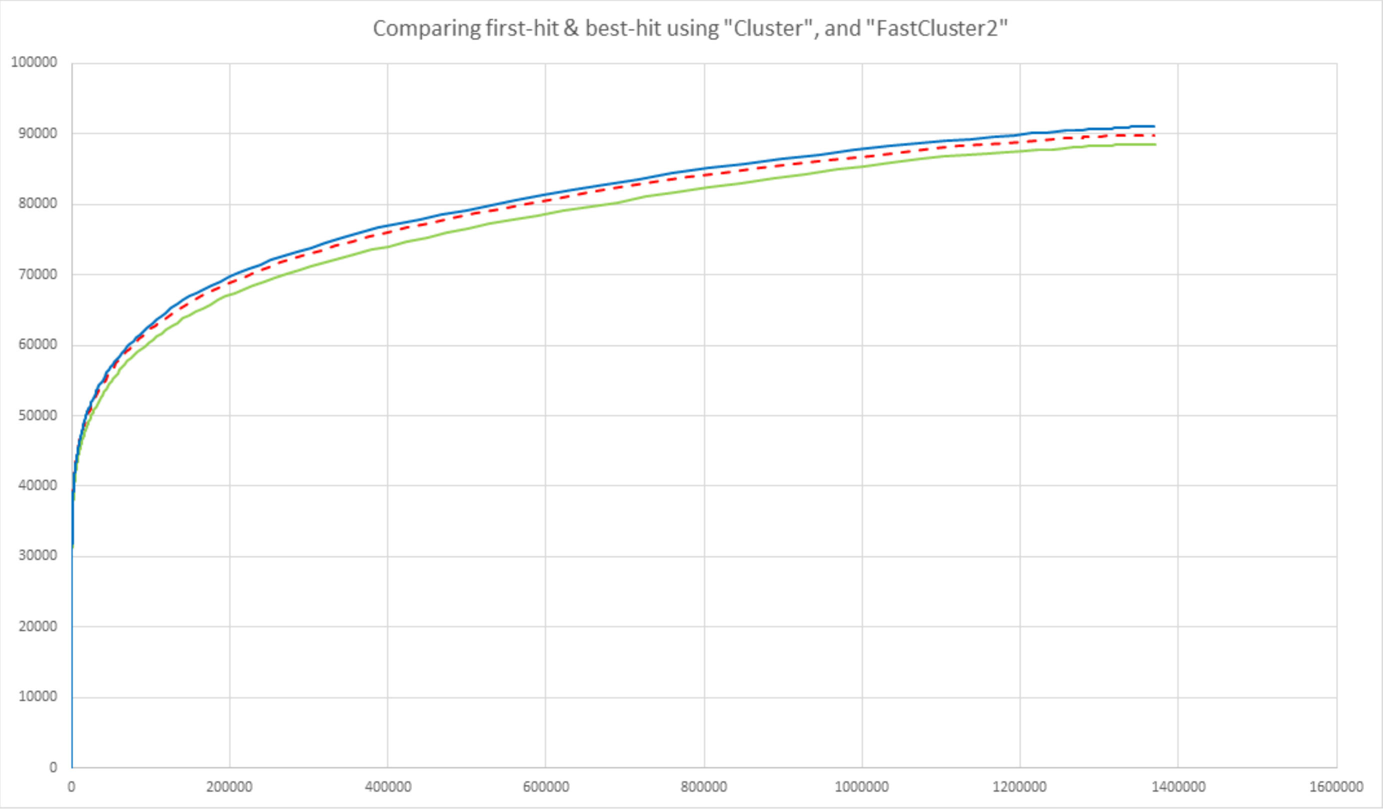
Description: filter by description

Gene Symbol: filter by gene symbol

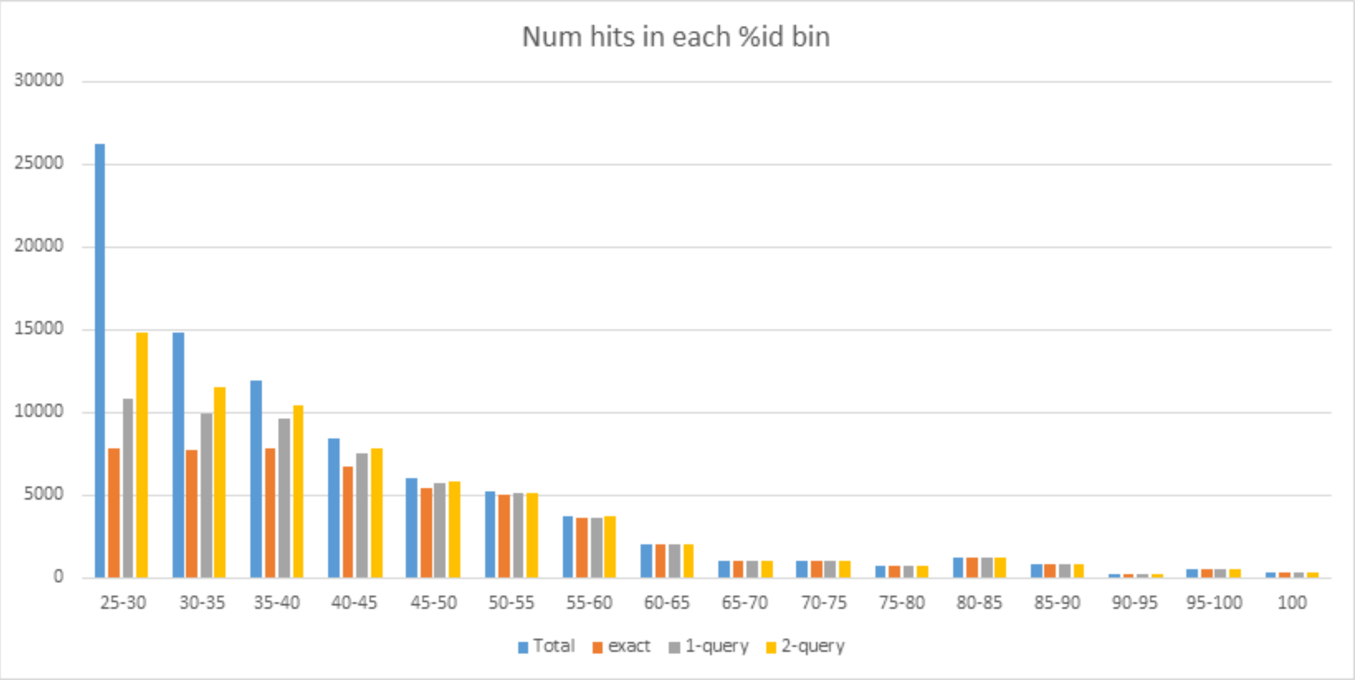
Identifier	Description	Organism	Alignment	Length	Identity	Actions
AHN56049	Wnt oncogene analog 2, isoform B	Drosophila melanogaster		325	42%	Fasta, Seqr, Links
NP_001286251	Wnt oncogene analog 2, isoform B	Drosophila melanogaster		325	42%	Fasta, Seqr, Links
CAA46001	Wnt-2 protein	Drosophila melanogaster		352	42%	Fasta, Seqr, Links
AAO24959	RE36604p	Drosophila melanogaster		352	42%	Fasta, Seqr, Domains
AAF58933	Wnt oncogene analog 2, isoform A	Drosophila melanogaster		352	42%	Fasta, Seqr, Links
NP_476810	Wnt oncogene analog 2, isoform A	Drosophila melanogaster		352	42%	Fasta, Seqr, Links
P28465	Protein Wnt-2	Drosophila melanogaster		352	42%	Fasta, Seqr, Links
AFH97153	FI20276p1	Drosophila melanogaster		352	42%	Fasta, Seqr, Links
AAY55697	IP02562p	Drosophila melanogaster		304	36%	Fasta, Seqr, Domains


filter stop-words
100000
95000
90000
85000
80000
75000
70000
65000
60000
55000
50000
0
200000
400000
600000
800000
1000000
1200000
1400000


reassign 0 vs reassign 8
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
200000
400000
600000
800000
1000000
1200000
1400000
1600000


Comparing first-hit & best-hit using "Cluster", and "FastCluster2"
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
200000
400000
600000
800000
1000000
1200000
1400000
1600000

Comparison to delta BLAST


Num hits in each %id bin
Total
exact
1-query
2-query