Genome OLAP:

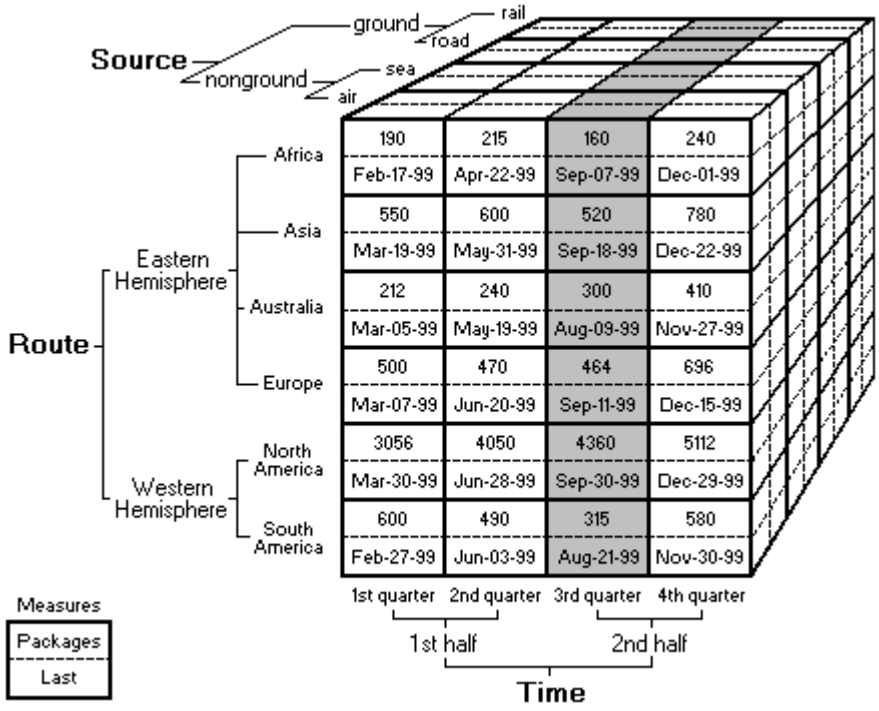
Online Tools to Mine Sequence Similarities Using Protein Annotations.

Corwin Joy

cjoy@bcm.tmc.edu

Human Genome Sequencing Center, Baylor College of Medicine

OLAP (an abbreviation for "Online Analysis and Processing") is a type of database technology that has long been used by the business community to analyze and interactively explore large financial data sets. The basic idea is that data sets are viewed as cubes with hierarchies along each axis. To navigate the cube, we specify an aggregation function to say how we want to aggregate summary information about groups of cells within the cube.



To apply this concept to the biology domain, let us begin by examining a table from the orthologous comparison between the Human and Chimpanzee genomes recently published in Nature:

Initial sequence of the chimpanzee genome and comparison with the human genome,

The Chimpanzee Sequencing and Analysis Consortium, Nature 437, 69-87, Sep. 1, 2005.

Table 5. GO categories with the highest divergence rates in hominids.

Next table | Previous table | Figures & Tables index

GO categories within 'biological process'	**Number of orthologues**	**Amino acid divergence**	K_A/K_S
Listed are the ten categories in the taxonomy biological process with the highest K_A/K_S ratios, which are not significant solely due to significant subcategories.			
GO:0007606 sensory perception of chemical stimulus	59	0.018	0.590
GO:0007608 perception of smell	41	0.018	0.521
GO:0006805 xenobiotic metabolism	40	0.013	0.432
GO:0006956 complement activation	22	0.013	0.428
GO:0042035 regulation of cytokine biosynthesis	20	0.011	0.402
GO:0007565 pregnancy	34	0.014	0.384
GO:0007338 fertilization	24	0.010	0.371
GO:0008632 apoptotic programme	36	0.010	0.358
GO:0007283 spermatogenesis	80	0.008	0.354
GO:0000075 cell cycle checkpoint	27	0.006	0.354

As published, this analysis is a flat table that shows amino acid divergence in just one dimension. However, the analysis could be easily extended by using OLAP to better present and understand these results. Specifically, GO has generalized hierarchies that describe **molecular function**, **biological process** and **cellular location.** We can use OLAP to browse divergence along each of these dimensions. Furthermore, the data set in this paper compares sequences between several different species so we can also incorporate a taxonomy dimension to see how similarity varies by species classification.

To see how an OLAP sequence browser works, we decided to start with a small data set rather than an entire genome. As a first example, we chose a set of related GPCR protein sequences. To the sequence data we added two annotations that we will use for browsing the sequences. The first annotation is taxonomy information about the species that the protein comes from. The second annotation is protein function. The test data is a series of 15 related rhodopsin and olfactory GPCRs from the GPCR database at gpcr.org.

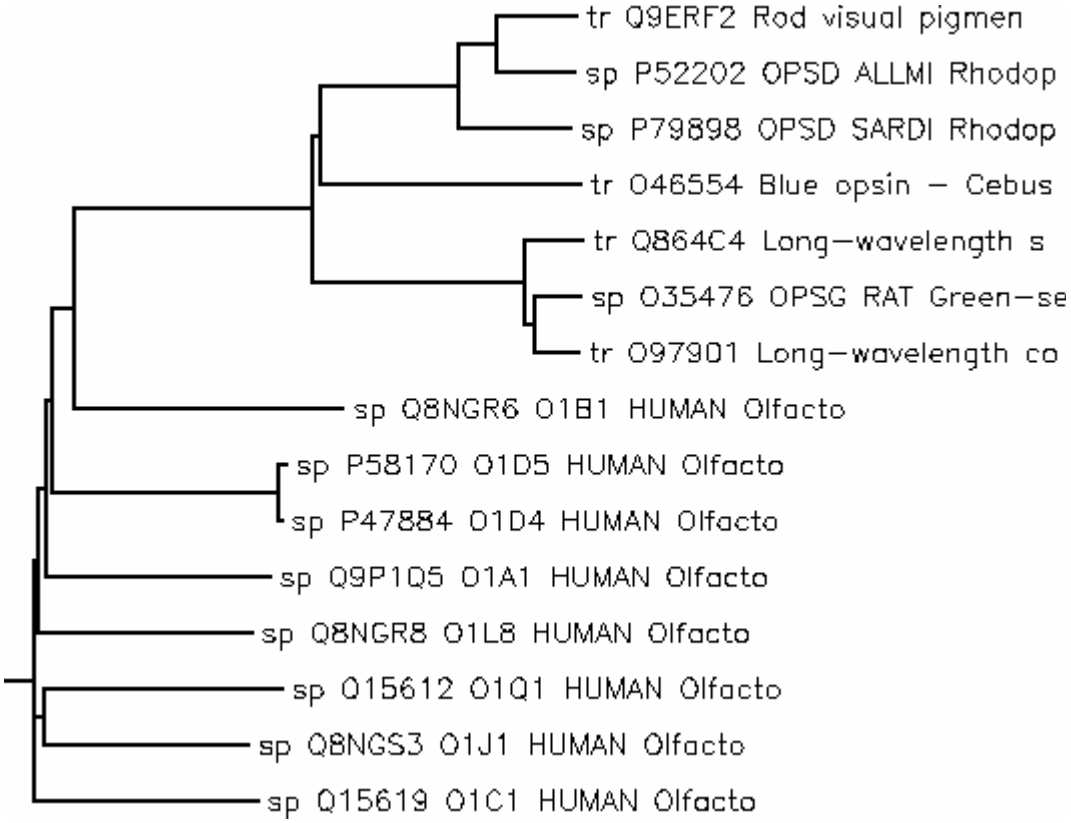
Olfactory sequences: Rhodopsin like Olfactory II family 1

http://www.gpcr.org/seq/001_005_001/001_005_001.html
Swissprot codes: O1A1_HUMAN, O1B1_HUMAN , O1C1_HUMAN, O1D4_HUMAN, O1D5_HUMAN, O1J1_HUMAN, O1Q1_HUMAN, O1L8_HUMAN

Comparison family: Rhodopsin Vertebrate

http://www.gpcr.org/seq/001_004_001/001_004_001.html
O46554, O97901, OPSD_ALLMI, OPSD_ALLMI, OPSD_SARDI, OPSG_RAT, Q864C4, Q9ERF2

For a small set of sequences like this, it is easy to view them via a rooted cladogram:



Can we obtain similar results to what is given by the cladogram using OLAP tools to browse this set of sequences interactively? To test this, we need a measure that can show us how "similar" a set of sequences is. Our requirements for this measure are:

1. It should be fast to calculate so that large sets of sequences can be browsed interactively.
2. It should work even when applied to distantly related sequences.

No single number can properly summarize all the relations in a group, but we can obtain a fast and approximate answer by looking at the work done by RC Edgar in developing the MUSCLE alignment tool, Edgar R.C., "**Local homology recognition and distance measures in linear time using compressed amino acid alphabets.",** *Nucleic Acids Res* 2004, **32**(1)**:**380-385. Here Edgar found that just counting the number of 3-mers that two protein sequences have in common gives a 94% correlation with the sequence identity score obtained from a full alignment. We therefore chose this as a reasonable starting point for quickly calculating sequence similarity. The measure that we report for a group of sequences represents the average 3-mer similarity score between all pairs in the group. We apply this measure to our set of test sequences via an OLAP cube.

Can an OLAP cube allow us to browse sequence similarities interactively?
Let us start by looking at the most aggregated view of the data, with all proteins grouped together:

Example Query Against GPCR Rhodopsin-Like Sequences

		Measures	
Taxonomy	ProteinFunction	SequenceSimilarity	SequenceCount
+All Organisms	+All Proteins	50.7809523809524	15.0

Slicer:

back to index

We can expand this by protein function dimension. Proteins within the same functional category have a higher similarity.

Example Query Against GPCR Rhodopsin-Like Sequences

		Measures	
Taxonomy	ProteinFunction	SequenceSimilarity	SequenceCount
+All Organisms	-All Proteins	50.7809523809524	15.0
	-GPCR	50.7809523809524	15.0
	-Class A Rhodopsin Like	50.7809523809524	15.0
	(Rhod)opsin	104.714285714286	7.0
	Olfactory	67.3928571428571	8.0

Slicer:

back to index

Expand by taxonomy dimension. Proteins within the same taxonomy class and functional category have an even higher similarity.

Example Query Against GPCR Rhodopsin-Like Sequences

		Measures	
Taxonomy	ProteinFunction	SequenceSimilarity	SequenceCount
−All Organisms	−All Proteins	50.7809523809524	15.0
	−GPCR	50.7809523809524	15.0
	−Class A Rhodopsin Like	50.7809523809524	15.0
	(Rhod)opsin	104.714285714286	7.0
	Olfactory	67.3928571428571	8.0
−Animalia	−All Proteins	50.7809523809524	15.0
	−GPCR	50.7809523809524	15.0
	−Class A Rhodopsin Like	50.7809523809524	15.0
	(Rhod)opsin	104.714285714286	7.0
	Olfactory	67.3928571428571	8.0
−Chordata	+All Proteins	50.7809523809524	15.0
+Actinopterygii	+All Proteins	0.0	1.0
+Archosauria	+All Proteins	0.0	1.0
+Mammalia	−All Proteins	50.6153846153846	13.0
	−GPCR	50.6153846153846	13.0
	−Class A Rhodopsin Like	50.6153846153846	13.0
	(Rhod)opsin	117.0	5.0
	Olfactory	67.3928571428571	8.0

Drill through to see the underlying mammalian sequence data for mammalian rhodopsin sequences:

Drill Through Table for SequenceSimilarity

SPECIES	GENUS	FAMILY	ORDER	CLASS	PHYLUM	KINGDOM	PROTEIN_SUBFAMILY_	PROTEIN_CLASS	PROTEIN_FAMILY	
Black Sea Dolphin	Delphindae	Odontoceti	Cetacea	Mammalia	Chordata	Animalia	(Rhod)opsin	Class A Rhodopsin Like	GPCR	MAQTWGPQRFAGGQF
Ehrenberg's mole rat	Spalacinae	Muridae	Rodentia	Mammalia	Chordata	Animalia	(Rhod)opsin	Class A Rhodopsin Like	GPCR	MNGTEGPNFYVPFSN
Harbor seal	Seal	Phocidae	Carnivora	Mammalia	Chordata	Animalia	(Rhod)opsin	Class A Rhodopsin Like	GPCR	MAQTWGLQRLADGRP
Rattus Norvecicus	Murinae	Muridae	Rodentia	Mammalia	Chordata	Animalia	(Rhod)opsin	Class A Rhodopsin Like	GPCR	MAQQLTGEQTLDHYE
Weeper Capuchin	Capuchin	Platyrrhini	Primates	Mammalia	Chordata	Animalia	(Rhod)opsin	Class A Rhodopsin Like	GPCR	MSKMSEEEEFYLFKNI

So, looking at the similarity scores that are obtained from our simple k-mer measure, this seems to correspond to what we would expect from the rooted cladogram, but these k-mers can be calculated much more quickly and give us a way to easily explore large sets of sequences using various classifications. Once we have found a grouping that has a high similarity score we are interested in, the tool makes it very easy to display the underlying sequences for more refined alignment, motif extraction, and comparison.

Extension to the iProClass Database:

To test this tool on a larger scale we decided to apply the technology to the iProClass database (http://pir.georgetown.edu/iproclass/). This database contains over 2 million protein sequences, annotated with GO classification, PIR superfamilies, motifs, protein domains etc. All of these annotations are potentially useful as dimensions for a cube. We chose two obvious ones, the GO protein function hierarchy and the taxonomy / lineage information for the sequence. By restricting our focus to sequences containing these annotations we came up with a total of 44,441 sequences for our cube. In this case, the initial similarity calculation for these 44k sequences took over 10 minutes. Fortunately, the OLAP browser allows us to pre-calculate similarities and store them in a table so that the response time of the viewer is near instantaneous. Again, an OLAP browser allows us to easily navigate this large data set.

Expand one level deep by GO classification and top level Taxonomy classification:

		Measures	
GO	Taxonomy	SequenceSimilarity	SequenceCount
−All Terms	−All Species	↓23	↓44,411
	+Viruses	↓35	↓1,634
	+cellular organisms	↓14	↓42,770
	+null	↓108	↓7
+obsolete_molecular_function	−All Species	↓31	↓1,263
	+Viruses	↓9	↓3
	+cellular organisms	↓6	↓1,260
	+null		
+obsolete_biological_process	−All Species	↓13	↓50
	+Viruses	↓	↓1
	+cellular organisms	↓8	↓49
	+null		
+obsolete_cellular_component	−All Species	↓31	↓29
	+Viruses		
	+cellular organisms	↓27	↓29
	+null		
+molecular_function	−All Species	↓34	↓46,349
	+Viruses	↓83	↓1,337
	+cellular organisms	↓32	↓45,005
	+null	↓108	↓7
+cellular_component	−All Species	↓35	↓13,293
	+Viruses	↓43	↓211
	+cellular organisms	↓26	↓13,082
	+null		
+biological_process	−All Species	↓34	↓135,891
	+Viruses	↓70	↓2,121
	+cellular organisms	↓23	↓133,770
	+null		

Expand by GO Hierarchy:

		Measures	
GO	Taxonomy	SequenceSimilarity	SequenceCount
−All Terms	+All Species	23	44,411
+obsolete_molecular_function	+All Species	31	1,263
+obsolete_biological_process	+All Species	13	50
+obsolete_cellular_component	+All Species	31	29
+molecular_function	+All Species	34	46,349
−cellular_component	+All Species	35	13,293
+extracellular region	+All Species	48	73
−cell	+All Species	37	6,681
+cell fraction	+All Species	63	7
−intracellular	+All Species	52	6,457
+nucleus	+All Species	44	228
+cytoplasm	+All Species	39	2,867
+cytoskeleton	+All Species	55	148
+fimbrium	+All Species	20	41
+thylakoid	+All Species		1
+light-harvesting complex	+All Species	8	4
+ribonucleoprotein complex	+All Species	9	26
+intracellular organelle	+All Species	36	3,142
+membrane	+All Species	85	175
+external encapsulating structure	+All Species	21	31
+cell projection	+All Species	21	11
+virion	+All Species	33	86
+extracellular matrix	+All Species	58	67
+organelle	+All Species	26	6,355
+protein complex	+All Species	8	31
+biological_process	+All Species	34	135,891

From any cell we can drill down to view the underlying data for that group of sequences:

Drill Through Table for SequenceCount

TERM ID (Key)	TERM ID	CSQID	TAX6	TAX5	TAX4	TAX3	TAX2	TAX1	
15.00	trans-hexaprenyltranstransferase activity	E69630+HEP2_BACSU	Bacillaceae	Bacillales	Bacilli	Firmicutes	Bacteria	cellular organisms	MLNIIRLLAESLPRISDGNENTDVWVNDMKFKM
15.00	trans-hexaprenyltranstransferase activity	HEP2_BACST	Bacillaceae	Bacillales	Bacilli	Firmicutes	Bacteria	cellular organisms	MKLKAMYSFLSDDLAAVEEELERAVQSEYGPL
15.00	trans-hexaprenyltranstransferase activity	C69630+HEP1_BACSU	Bacillaceae	Bacillales	Bacilli	Firmicutes	Bacteria	cellular organisms	MQDIYGTLANLNTKLKQKLSHPYLAKHISAPKI
39.00	adenine deaminase activity	H69279+ADEC_ARCFU	Archaeoglobaceae	Archaeoglobales	Archaeoglobi	Euryarchaeota	Archaea	cellular organisms	MSSPTADVEKLRRIIEVARGDRRADFVVKNAQ
39.00	adenine deaminase activity	H69279+ADEC_ARCFU	Archaeoglobaceae	Archaeoglobales	Archaeoglobi	Euryarchaeota	Archaea	cellular organisms	MSSPTADVEKLRRIIEVARGDRRADFVVKNAQ
39.00	adenine deaminase activity	C70253+ADEC_BORBU	Spirochaetaceae	Spirochaetales	Spirochaetes (class)	Spirochaetes	Bacteria	cellular organisms	MDLFKIEANYIDIFNKEIYPASIAIANGHIASIEI
39.00	adenine deaminase activity	C70253+ADEC_BORBU	Spirochaetaceae	Spirochaetales	Spirochaetes (class)	Spirochaetes	Bacteria	cellular organisms	MDLFKIEANYIDIFNKEIYPASIAIANGHIASIEI
39.00	adenine deaminase activity	F69215+ADEC_METTH	Methanobacteriaceae	Methanobacteriales	Methanobacteria	Euryarchaeota	Archaea	cellular organisms	MISGNILNVFTGDIYPAEIEVAGGRVRCVRSIS
39.00	adenine deaminase activity	F69215+ADEC_METTH	Methanobacteriaceae	Methanobacteriales	Methanobacteria	Euryarchaeota	Archaea	cellular organisms	MISGNILNVFTGDIYPAEIEVAGGRVRCVRSIS
51.00	Rieske iron-sulfur protein	D70784+QCRA_MYCTU	Actinomycetales	Actinobacteridae	Actinobacteria (class)	Actinobacteria	Bacteria	cellular organisms	MSRADDDAVGVPPTCGGRSDEEERRIVPGPNP

Page 1/5,890 Goto Page 1

Conclusion:

In conclusion, OLAP seems to have potential as a way of interactively browsing large genomic data sets and annotations. The key to applying this technology to genomic data sets is to recognize that we can aggregate sets of genomic sequences just like we can aggregate sets of numbers. Useful aggregation functions might include sequence similarity, most conserved motifs, most prevalent amino-acids and so on. We can use sequence annotations as hierarchies similar to what we have shown here, and allow biologists to browse the genome to explore regions of high similarity (or divergence) within species, by gene function, or across various gene families.