



Vertebrates, flies and worms: Protein domain usage compared

Introduction

The vertebrate genome has increased its complexity both by evolving new protein domains, and more importantly, assembling protein domains into unique protein architectures. By combining a small number of protein domains in a large number of different ways, a tremendous variation in both protein structure and function is possible. The fact that vertebrates have created more unique protein architectures from a set of protein domains has allowed development of a more complex organism without a large increase in genome size or gene number.





New Protein Domains

A small part of the complexity of the vertebrate genome is due to invention of new protein domains.

Protein Domain

A protein domain is a part of a protein with a specific structure and function. Protein domains can be thought of as “building blocks” of entire proteins.

Examples of protein domains

-  Binds hormones
-  Binds cytokines
-  Kinase (adds a phosphate group)
-  Phosphatase (removes a phosphate group)

Molecules that bind hormones and cytokines are receptors, and are important in allowing a cell to recognize and respond to a signal from elsewhere in the body.

Addition and removal of phosphates is a way that cells use to activate or inactivate different enzymes thus turning on or off different cellular processes.

7% of the protein domains found in vertebrates are vertebrate specific.

Most protein domains came from a shared ancestor common to all animals.

Of the new domains found only in vertebrates, most are involved in immunity or the nervous system.

New Protein Architectures

Much of the complexity of the vertebrate genome is due to creation of new protein architectures.

Protein architecture

Protein architecture is the linear arrangement of domains within a protein. Protein architecture determines the structure and function of the entire protein, including all its domains.

Example – different organisms may have combined the same protein domains in differing numbers of unique combinations. In the example below Organisms A and B have the same protein domains (building blocks). Organism B however, has assembled its domains in a larger number of unique ways, leading to increased diversity of functions it is able to carry out.

Organism A



A protein that binds hormones and then adds a phosphate group



A protein that binds cytokines and then removes a phosphate group

Organism B



A protein that binds hormones and then adds a phosphate group



A protein that binds cytokines and then removes a phosphate group



A protein that binds cytokines and then adds a phosphate group



A protein that binds a hormone and then removes a phosphate group

You can see that organism B combined these protein domains in more ways, allowing it to carry out more functions.

Vertebrates have shuffled pre-existing protein domains into a more complex set of protein architectures thus allowing more complex functions.

Humans have 1.8 times as many protein architectures as fly or worm and 5.8 times as many as yeast.

The increase in protein architectures is particularly evident in the development of new extracellular and transmembrane architectures.

Activity

In this activity, we will search the number of different proteins in which various domains occur. We will search the genomes of several different organisms including, human, mouse, fly, and nematodes (a type of worm).

Protein domains we will work with

- **Immunoglobulin domain** – Domains involved in the recognition of specific antigens and signals.
- **C2H2-type zinc finger** - A domain often found in transcription factors. These domains interact with DNA.
- **Eukaryotic protein kinase** – Adds phosphate groups to proteins (usually activating them)
- **P-loop motif** – found in DNA binding proteins. This domain is involved in binding, recognition and regulation of activity of different genes.
- **Reverse transcriptase** – Makes a DNA molecule from an RNA template.
- **Trypsin-like serine protease** – Removes phosphate groups from serine (usually inactivating the protein containing the serine).
- **Ankyrin repeat** – tandemly repeated domain of 33 amino acids. This domain occurs in a large number of functionally diverse proteins. It has a conserved L-shaped fold structure. The ankyrin repeat functions in protein-protein interactions.
- **RING finger** – are involved in protein interactions.
- **Leucine-rich repeat** - Leucine Rich Repeats are short sequence motifs present in a number of proteins with diverse functions and cellular locations. Leucine Rich Repeats are often flanked by cysteine rich domains.
- **Collagen triple helix repeat** – This domain is a repeat of 3 amino acids, glycine, proline, hydroxyproline. When repeated in a sequence, this repeat will form a helix. This repeat is found in collagen and sometimes other proteins.

Activity

1. Go to the NCBI website
<http://www.ncbi.nih.gov/>

2. Click on Human Genome Resources in the right column

The screenshot shows the NCBI homepage. At the top, it says "National Center for Biotechnology Information" with subtext "National Library of Medicine" and "National Institutes of Health". Below this is a navigation bar with links for PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, and Structure. A search bar is present with "Entrez" selected in the dropdown and a "Go" button. On the left, there is a "SITE MAP" section with links to "About NCBI", "GenBank", and "Literature databases". In the center, there is a "What does NCBI do?" section with a paragraph of text and a "PubMed Central" section with bullet points. On the right, there is a "Hot Spots" column with several links, including "Human genome resources", which is highlighted by a red arrow.

3. Type your domain name (shown on the previous page) into the box next to locus link and click on "go." (Note, if you are not on the human genome page, be sure to select "human" from the organism box to limit your search to the human genome – not shown on this particular screen which is from the human genome page).

The screenshot shows the "Human Genome Resources" page. At the top, it says "NCBI Home > Genomic Biology > Homo sapiens". Below this is a search bar with "LocusLink" selected in the dropdown and a "Go" button. On the left, there is a "Browse Your Genome" section with a "show" dropdown set to "Genes" and a grid of chromosome icons numbered 1 through 22, plus X and Y. Below this is a link to "The NCBI Handbook". In the center, there is a paragraph of text about the challenge of piecing together genomic data. On the right, there is a "THE GENOMIC SEQUENCE" section with a "1953" icon and text about the reference DNA sequence of Homo sapiens, and a "BLAST the Genome" section with text about comparing sequences and a "Clone Registry" link. A red arrow points from the search bar in the previous screenshot to the search bar in this screenshot.

- A list of the proteins in which this domain is found will come up
Note the number of proteins in which this domain is found (in this case 75) in the table on the next page.
- Below, list at least two different proteins that contain your protein in humans. The names of proteins are often written out in the description column on this first page

NCBI LocusLink

PubMed Entrez BLAST OMIM Map Viewer Taxonomy Structure

Search: LocusLink Display: Brief Organism: Human

Query: P-loop motif Go Clear

View Loci Save Loci

75 loci found This is page 1 of 2

LocusID	Org	Symbol	Description	Position	Links
<input type="checkbox"/> 89941	Hs	ARHT2	ras homolog gene family, member T2	16p13.3	P R G P H U V
<input type="checkbox"/> 116986	Hs	CENTG1	centaurin, gamma 1	12q13.2	P O P H U V
<input type="checkbox"/> 116987	Hs	CENTG2	centaurin, gamma 2		P R G P H U V
<input type="checkbox"/> 120892	Hs	DKFZp434H2111	hypothetical protein DKFZp434H2111	12q12	R G P H U V

- Optional advanced question - What are the functions of these proteins? You will need to click on the locus ID number or one of the additional links such as PubMed or OMIM to find the function of a protein. You may want to choose proteins for which it is easy to find a function (not all proteins listed will have a clearly defined function). Refer to the NCBI website instructions for a more detailed description of how to find protein function.

- Then use the organism box in the top gray bar to change your search to mouse, fly or nematode (a type of worm). Click on go.
- Repeat until you have filled out the table for your domain.
- Complete the table with data from the entire classrooms.

	Human	Mouse	Fly	Nematode
Immunoglobulin domain				
C2H2-type zinc finger				
Eukaryotic protein kinase				
P-loop motif				
Reverse transcriptase				
Trypsin-like serine protease				
Ankyrin repeat				
RING finger				
Leucine-rich repeat				
Collagen triple helix repeat				

Teacher answers

These were the answers obtained when the website was searched in June 2003. These answers will continually change over time, but the trend towards wider domain usage (larger numbers) in vertebrates should remain.

	Human	Mouse	Fly	Nematode
Immunoglobulin domain	330	551	39	29
C2H2-type zinc finger	22	1	0	6
Eukaryotic protein kinase	213	40	2	24
P-loop motif	75	82	7	0
Reverse transcriptase	38	77	0	113
Trypsin-like serine protease	116	146	25	0
Ankyrin repeat	129	117	10	13
RING finger	277	293	22	63
Leucine-rich repeat	93	98	5	4
Collagen triple helix repeat	22	30	0	1