Available online at www.sciencedirect.com
ScienceDirect
Developmental Biology 300 (2006) 485–495
DEVELOPMENTAL BIOLOGY
www.elsevier.com/locate/ydbio

The *S. purpuratus* genome: A comparative perspective

Stefan C. Materna, Kevin Berney, R. Andrew Cameron *

Division of Biology, m/c 139-74, California Institute of Technology, 1200 East California Blvd., Pasadena, CA 91125, USA

Received for publication 31 May 2006; revised 15 September 2006; accepted 19 September 2006
Available online 26 September 2006

Abstract

The predicted gene models derived from the sea urchin genome were compared to the gene catalogs derived from other completed genomes. The models were categorized by their best match to conserved protein domains. Identification of potential orthologs and assignment of sea urchin gene models to groups of homologous genes was accomplished by BLAST alignment and through the use of a clustering algorithm. For the first time, an overview of the sea urchin genetic toolkit emerges and by extension a more precise view of the features shared among the gene catalogs that characterize the super-clades of animals: metazoans, bilaterians, chordate and non-chordate deuterostomes, ecdysozoan and lophotrochozoan protostomes. About one third of the 40 most prevalent domains in the sea urchin gene models are not as abundant in the other genomes and thus constitute expansions that are specific at least to sea urchins if not to all echinoderms. A number of homologous groups of genes previously restricted to vertebrates have sea urchin representatives thus expanding the deuterostome complement. Obversely, the absence of representatives in the sea urchin confirms a number of chordate specific inventions. The specific complement of genes in the sea urchin genome results largely from minor expansions and contractions of existing families already found in the common metazoan "toolkit" of genes. However, several striking expansions shed light on how the sea urchin lives and develops.

Keywords: Metazoan; Deuterostome; Bilaterian; Gene catalog; Protein domain

Introduction

In the six-kingdom scheme to organize all life forms revised by Cavalier-Smith (1998), the animal kingdom is divided into three subkingdoms the most diverse of which is the Bilateria (Fig. 1). This division contains most of what we normally think of as the large animals. The bilaterian lineage arose in the late Neoproterozoic and predates the Cambrian explosion, 540 million years ago (Adoutte et al., 2000; Balavoine and Adoutte, 1998). One of several monophyletic groups that shares a common ancestor with the bilaterians is Cnidaria (anemones and jellies), a phylum that is currently the best outgroup comparison (reviewed in Eernisse and Peterson, 2004).

The bilaterians are divided into Protostomia and Deuterostomia, a naming convention that is just 100 years old (Grobben, 1908). The characters used to delineate this division were embryological: the pattern of cleavage, the origins of the digestive tract and the manner in which the mesoderm formed (reviewed in Hyman, 1954). The form of the larvae was another important character (reviewed in Nielsen, 1995). Although much controversy has accompanied the assignment of individual phyla to these super-clades over the century, a combination of molecular and morphological evidence strongly supports the scheme shown in Fig. 1 (Halanych, 2004; Peterson and Eernisse, 2001). The Protostomia are divided into ecdysozoan and lophotrochozoan super-clades on the basis of molecular characters derived from ribosomal RNA or mitochondrial DNA (Halanych et al., 1995; Aguinaldo et al., 1997). The deuterostomes are divided into two super-clades, non-chordates, including sea urchins and hemichordates, and chordates, including cephalochordates, urochordates and vertebrates (Castresana et al., 1998; Turbeville et al., 1994; Wada and Satoh, 1994). The phylum Echinodermata has five classes: the sea urchins (Echinoidea), the sea stars (Asteroida), sea cucumbers (Holothuroidea), brittle stars (Ophiuroidea) and sea lilies (Crinoidea). Thus the sea urchins are invertebrates in the lineage leading to the vertebrates and humans.

* Corresponding author. Fax: +1 626 795 3382.
E-mail address: acameron@caltech.edu (R.A. Cameron).

0012-1606/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ydbio.2006.09.033

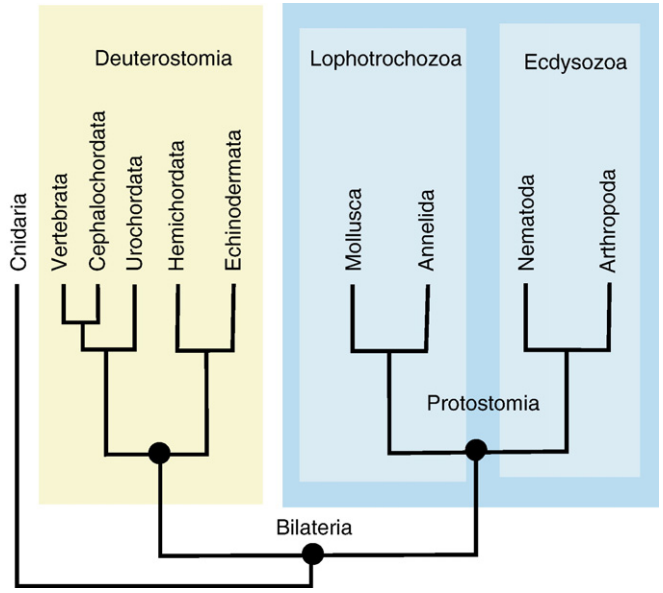

Fig. 1. A simplified view of bilaterian phylogeny for the taxa discussed in the text and based on both molecular and morphological data. The outgroup for the bilaterians is the Cnidaria which genomic information shows to be more closely related to the bilaterians than the Ctenophores. The major branches of the Bilateria, Deuterostomia and Protostomia are indicated as closed circles on the tree. The super-clades are indicated above the tree.

Our purpose here is to examine the gene complement of the newly available genome sequence of the sea urchin, *Strongylocentrotus purpuratus*, in comparison to that of other animals. Since the sea urchin is the perfectly positioned outgroup to the chordates, for the first time an overview emerges of the shared features of the gene catalogs that characterize the super-clades of the metazoans.

Materials and methods

Sequence databases

In order to compare the sea urchin gene set to that of other species, we obtained the set of GLEAN gene predictions based on the Spur v0.5 assembly from Baylor Sequencing Center (GenBank accession number AAGJ01000000; also referred to as NCBI build v1.1). This assembly matched about 84% of the ESTs for this species and demonstrated a redundancy level of 13% (The Sea Urchin Genome Sequencing Consortium, in press). The total number of predicted gene models in the set is 28,944 (Sodergren et al., 2006). Given the accuracy of the assembly, the redundant fraction of the gene set could be as much as 13% though it is probably less. It is likely that the overestimate is mainly due to haplotype differences that could not be resolved sufficiently during assembly. Since it is not currently known which among these is a genuine duplication, the analysis was carried out with the full GLEAN set.

The mouse and human protein sets were obtained from the International Protein Index (IPI, Kersey et al., 2004). This is a regularly updated collection of sequences derived from entries in various public repositories. The *Gallus gallus* (chicken) protein set was downloaded from Ensembl (http://www.ensembl.org/index.html). The *Drosophila melanogaster* (fruit fly) protein set was generated from the Release 3 assembly (Celniker et al., 2002; Misra et al., 2002). The *Ciona intestinalis* (ascidian) data set was created at the Joint Genome Institute (Dehal et al., 2002). The set of *Caenorhabditis elegans* (Nematode worm) proteins was obtained from wormbase (WS130, October 2004). The above protein sets are all the result of multiple iterations of the original protein predictions for these genomes and thus are well characterized.

The *Nematostella vectensis* (anemone) proteome was obtained from the Joint Genome Institute. It is based on version 1.0 of the *N. vectensis* genome. This set comprises 27,273 protein sequences. Based on the size of the *N. vectensis* genome assembly, redundancy is fairly minimal. However, this sequence collection includes many low-complexity peptides that presumably were not eliminated during the prediction process. We inspected BLAST (Altschul et al., 1997) searches involving *N. vectensis* for hits to low complexity sequences, which generally do not pass the given '*E*-value' thresholds. Thus, because of the early state of the *N. vectensis* genome, the results involving this organism have to be viewed as preliminary. The URLs for sequence downloads are listed in Supplementary Table 1.

The databases of mouse, chicken, fruit fly and worm contain multiple isoforms of proteins. These inflate the proteome of at least mouse and human significantly. Since we are interested in the abundance of domains encoded in the genome and not the proteome, we generated a non-redundant set of proteins that eliminates splice forms keeping only the longest corresponding protein for each gene. The OrthoMCL clustering algorithm (see below) is also sensitive to splice forms. If they exist in more than one species, they may end up in different clusters. We used the non-redundant protein sets for all our analyses (Supplementary Table 2).

IPRSCAN analysis

In order to assess the abundance of known protein domains in the sea urchin, the translations of GLEAN gene predictions were matched to Hidden Markov models (Eddy, 1998) by the IPRSCAN software from EBI (InterPro Consortium, 2001). The models included in this search were taken from the two most commonly utilized databases, PFAM (Sonnhammer et al., 1998) and SMART (Letunic et al., 2006) and complemented by several smaller databases (BlastProDom, Coil, Panther, PIR, Tigr, ProfileScan, ScanRegExp, Seg, Superfamily) which are part of InterPro (http://www.ebi.ac.uk/interpro/). To allow for consistent comparison, we also performed this analysis locally on the non-redundant gene sets from the genomes of mouse, fly, worm and ascidian.

Our aim is to identify signature domains for each protein that will inform us about its function. Therefore, multidomain proteins were classified according to (a) the domain with most occurrences in it or (b) the domain with the most significant '*E*-value' when the same number of domains was present. We assume that the most prevalent domain in a given protein is the one that best characterizes it.

Identification of orthologous groups

We performed clustering of an all versus all BLAST of *S. purpuratus*, *Mus musculus*, *G. gallus*, *D. melanogaster* and *N. vectensis* using the OrthoMCL software (Li et al., 2003). OrthoMCL clusters related sequences from different species which are potential orthologues and tries to distinguish in-paralogues from out-paralogues (Li et al., 2003). Our goal in conducting this analysis is to provide a rough overview of relationships at the genomic level. Clustering based on BLAST results cannot substitute for more in depth phylogenetic analysis. The '*E*-value' cutoff used for this analysis was 10^{-5}. The OrthoMCL inflation parameter was set to 1.5.

Results and discussion

IPRSCAN analysis

In order to computationally describe the variety of proteins, it is helpful to identify them by the protein domains they contain. Often these are characteristic of specific processes and allow conclusions about their biology. Thus, a genome-wide domain search gives us a broad view of the functions that are encoded in the genes. It allows for a quick comparison of different organisms' genomes and provides us with the unique opportunity to recognize common features and species- or clade-specific adaptations.