Origin of multicellular eukaryotes – insights from proteome comparisons
L Aravind* and G Subramanian†

The complete genomes of the yeast Saccharomyces cerevisiae and the nematode worm Caenorhabditis elegans have recently become available allowing the comparison of the complete protein sets of a unicellular and multicellular eukaryote for the first time. These comparisons reveal some striking trends in terms of expansions or extensive shuffling of specific domains that are involved in regulatory functions and signaling. Similar comparisons with the available sequence data from the plant Arabidopsis thaliana produce consistent results. These observations have provided useful insights regarding the origin of multicellular organisms.

Introduction
Ever since the discovery of microscopic lifeforms by van Leeuwenhoek in 1675, the tenous line dividing single-celled and multicellular life forms has been appreciated. Organisms like Volvox, slime molds, fungi, choanoflagellates and sponges provide a gradation of organizational levels that suggests a possible evolutionary scenario for the origin of multicellular forms from the unicellular versions. Even within these simple multicellular forms a basic functional differentiation is seen, with some cells playing a reproductive role and others playing a ‘structural’ or ‘nutritional’ role. Thus, the issue of explaining the loss of reproductive abilities of certain cells for the sake of others within these multicellular units remains.

In the basic framework of the Darwinian evolutionary paradigm, it appears likely that, in a colony arising from the clonal expansion of a single cell, kin selection could function to result in a differential multicellular form. A set of cells in a colony would differentiate and thereby sacrifice their reproductive ability for that of their clones (kin) if this behavior favors a net increase in fitness. Thus, the fixation of multicellularity would proceed with the selection of genes that favor the increase of fitness of that state relative to the individually reproducing unicellular counterparts.

What are the genes that could favor the existence of the multicellular state? The first glimpses of this genetic machinery have emerged in the past year with the first determination of the complete sequence of a multicellular eukaryote — the nematode Caenorhabditis elegans [1••]. Considerably, considerable advances have been made in terms of sequencing the genome of another multicellular eukaryote — the plant Arabidopsis thaliana [2••]. These multicellular eukaryotes develop from a single cell (the zygote) that replicates to give rise to a spatially structured body with a number of differentiated cell types amongst which only a specialized subset participates in reproduction [3]. This can be contrasted with the yeast Saccharomyces cerevisiae which shows temporal differentiation in terms of the various gene expression states arising in response to environmental conditions such as presence of nutrients or phenotype and starvation but with very little spatial complexity or specialization [4•]. All these three organisms — the yeast, the animal and the plant — belong to the crown group of eukarya and share a relatively recent common ancestor with respect to the rest of the eukaryotes [5,6]. Thus in the comparison of multicellular genomes to that of yeast, one may expect the molecular basis for multicellularity to stand out over the general background of phylogenetic affinity seen between these organisms. We briefly review here the highlights of the findings obtained from the comparison of the predicted protein sequences of S. cerevisiae with that of C. elegans and A. thaliana and consider them in the context of the evolution of the differentiated state.

Taking snapshots of the worm and yeast proteomes
In silico studies on the predicted proteomes of organisms have been vastly aided by the recent development of sensitive sequence/profile/analysis methods that allow objective detection and statistical evaluation of subtle sequence similarities. Amongst these methods, PSI-BLAST (position-specific iterated basic local alignment search tool) that uses a position specific weight matrix derived from the alignments obtained from the primary gapped BLAST search to iteratively search the database is particularly powerful in terms of sensitivity and speed [7]. The other effective technique implemented in programs like HMMER searches databases with user-supplied multiple alignments represented as a hidden Markov model [8••]. As a result of such searches, the protein domain families in a proteome can be comprehensively enumerated and trends in their distributions can be characterized.

Regulatory components of cellular processes such as signal transduction, transcription, cell-cell interactions and assembly of multimeric cellular components are most likely targets of selective forces in generating prominent differences between related organisms. Hence, to home in on changes that accompanied the origin of multicellular
states, a comparison between the yeast and nematode proteomes was conducted using domains most frequently found in regulatory proteins. Using databases of such regulatory domains (such as SMART [simple modular architecture research tool] [9••]) profiles were constructed for each of the domain families and the numbers of proteins containing them in each of the two proteomes were determined [10••]. This revealed that there were several substantially larger regulatory domain families in C. elegans that were entirely absent in yeast and even in the case of those that were shared, C. elegans often possessed larger numbers of proteins with these domains. To account for the ~3-fold increase in the number of proteins encoded by C. elegans with respect to S. cerevisiae, the above protein counts were normalized by the number of proteins present in each of the proteomes. Even in terms of normalized counts, specific domain families in C. elegans show a prominent expansion relative to their numbers in yeast (Figure 1a). An examination of these expanded families provided some immediate clues regarding animal adaptations that include multicellularity.

**Lineage-specific domain expansions**

The large C. elegans-specific domain families that were entirely lacking in yeast included domains mediating extracellular adhesion or those binding DNA and predicted to regulate transcription. These novel extracellular adhesion modules often contain cysteines that form disulfide bonds such as the
epidermal growth factor (EGF) and C-type lectin domains (Figure 1a). The novel DNA-binding domains belong to sev-
eral distinct families that include the metal chelating forms such as the nuclear hormone receptor family [11•], the T-box [12•] and the paired box [13•]. In the case of families that were shared by these organisms but expanded in *C. elegans*, similar trends were observed with certain transcription regu-
lation related domain families such as the homeodomain and the POZ domain displaying prominent expansions (Figure 1a) [14••,15•]. Further, modules such as the immunoglobulin domain, the von Willebrand A domain (TFIIH subunit p44), the fibronectin III (CHS5) domains and the leucine repeat that were found in intracellular contexts in the yeast [16•,17] were expanded in the worms and predominantly present in the extracellular segments of proteins. The domains participating in phosphotyrosine sig-
nal transduction such as the tyrosine kinases, SH2 domains, the phosphotyrosine phosphatases and the phosphotyrosine binding domain were strikingly expanded in *C. elegans* [18,19]. Smaller expansions were observed in the F-box, GTPase and receptor-like kinase domains, cyclic nucleotide cyclases, calcium binding EF-hand, cation channels and the F-box that targets the ubiqui-
tin-mediated degradation pathway [20•,21•,22,23•].

The invention and proliferation of specific transcription factor families in animals correlates well with the increase in drastically different cell types and spatial complexity that is achieved through a developmental process. The increase and recruitment of extracellular-interaction domains not only correlates with the need for adhesion in multicellular organisms but also with the need for specific intercellular contacts required for tissue development. Related to this intercellular communica-
tion is the rise of intracellular signaling domains that facilitate response to external stimuli. Of particular inter-
est in this group is the proliferation of the PDZ domains that bind to carboxy-terminal peptides of membrane pro-
teins and thereby act as adaptors for intracellular cascades (Figure 1a) [24]. Thus, it appears that the rise of animal multicellularity can be correlated to the selec-
tion of protein families that allowed differentiation (the transcription factors) and cellular communication (the signaling and adhesion specific domains). In contrast to these expansions, the only one unique to yeast is that of the C6 zinc cluster DNA binding domain.

**Trends in domains with lower relative numerical differences between the proteomes**

There are other families the normalized counts of which show a comparable value in the nematode and yeast. A careful analysis of these families shows that, even here, diverse trends are noticeable in the domain architec-
tures of these proteins. To clarify this point, the regulators of small GTPase signaling namely the GTPase activ-
ating proteins (GEFs) and GAPs (guanine nucleotide exchange factors) for Rho, Ras and Arf family of proteins were inves-
tigated [25•,26,27] (Figure 2a). The absolute numbers of these proteins and the domains found in them showed an expansion in *C. elegans*, whereas the total number of domains found per protein was comparable in the two organisms (Figure 2b). The number of different types of domains found in these proteins, however, showed a two-
fold increase in *C. elegans* with very few domain architectures being conserved between the two taxa (Figure 2). This suggests that there have been several novel domain juxtapositions during the origin of multicol-
ergous animals. A careful examination of their domain architectures and sequence similarity reveals that almost none of them are common to the two proteomes with a greater number of finger domains per protein seen in *C. ele-
gans*, suggesting that binding to larger regulatory elements in the multicellular genomes selected for the independent expansion of multidomain C2H2 fingers in animals.

The process of differentiation in animals involves asymmetric division and lateral inhibition to designate certain...
Figure 2

(a) Protein count, domain count, domain/protein, type of domains

(b) Common proteins: C. elegans, S. cerevisiae

Origin of multicellular eukaryotes — insights from proteome comparisons Aravind and Subramanian 691
members of a cellular field for a given differentiation state [29]. A key signaling pathway that participates in this process is the Notch pathway that includes the receptor protein Notch, its ligands and modifier, as well as intracellular proteins such as the transcription factor Su(H)/CBF1 (suppressor of hairless/C promoter binding factor) that transmits the signal to the nucleus [30]. It is seen that whereas the receptor Notch and its ligand may be unique to animals, the extracellular modifier (the glycosyltransferanse Fringe [31]) and the intracellular components (DNA-binding domain Su(H) and chromatinic ATPase Strawberry Notch) are most likely ancient components as their homologs can be found in plants and Schizosaccharomyces (L. Aravind, unpublished data). This suggests that the pathway did not arise all at once with the rise of animal multicellularity but rather by recruiting pre-existing proteins and adding a few new inventions. Similarly, recruitment followed by divergence could have given rise to the novel domain inventions in animals such as the Paired domain HTHs (helix turn helices) from transposase HTH domains [32], the Hedgehog like proteins from selfish elements like inteins [33] and the NHR (nuclear hormone receptor) zinc finger from the Lim domains [34].

Opposed to the prominent differences in the regulatory proteins is the relative conservation of domains found in certain general chromatinic factors such as the SWI/SNF ATPases and the cell cycle regulatory proteins [35]. Nevertheless, detailed investigation of these systems reveals some rather interesting distinctions between unicellular fungi and animals. The structural chromosomal proteins that are involved in positive and negative regulation of chromatin structure show a detectable increase in diversity of domain architectures and domain numbers per protein in animals. Selection for the establishment of multiple distinct chromatin states in the multicellular organisms would have resulted in this diversification of the structural components while maintaining a relatively constant set of catalytic activities in the form of the ATPases. The core cell cycle components such as the MCMs (minichromosome maintenance proteins) and the CDKs (cyclin-dependent kinases) are universally found in the common ancestor of the crown group but the general lack of conservation of domain architectures suggests that they were independently recruited to distinct subsets of cells in plants and animals to the exclusion of the yeast (L. Aravind, unpublished data). These proteins may have been selected for in multicellular eukaryotes specifically to regulate proliferation as has been suggested by studies on these proteins in animals [36,37].

The story of the weed – Arabidopsis thaliana

These comparisons of the complete proteomes of C. elegans and S. cerevisiae suggest that specific domain families may have undergone expansion and combination with other domains in response to the forces selecting for the maintenance of multicellularity. The availability of significant amounts of A. thaliana sequence data [28] allows a similar independent comparison between a distinct multicellular organism and yeast to test if there are any general trends in the evolution of multicellularity. The results of such a comparison are very consistent with the C. elegans/S. cerevisiae yeast comparison (Figure 1b). The plants show the invention and expansion of two novel transcription factor families, AP2 (Apetala-2 domain) and viviparous1 [38,39], and also a proliferation of shared transcription factors such as MYB, bHLH and the MADS domain (Figure 1b).

The plants also show the recruitment of distinct extracellular adhesion modules such as the bulb lectin and an EGF-like cysteine rich module and LRR just as in animals. In terms of increasing in the counts of intracellular signaling domains, the protein kinases, PP2C phosphatases, EF hands and the F-boxes figure prominently. Interestingly, there is no expansion of the PDZ domain suggesting that this feature may be linked to the animal-specific development of neural tissue and associated proliferation of the cation channels. The plants also draw extensively from their cyanobacterial inheritance to develop bacterial type signaling systems in the form of multi-ple two-component system proteins — kinases and receiver domain proteins [40]. Although similarities in the expansion trends are observed between animals and plants, even the common domain expansions such as that of the homeodomain, the F-box, LRR and the protein kinases appear to have occurred independently. In all these cases, analogous rather than orthologous domain organizations are seen, for example animal receptor tyrosine kinases compared with the plant receptor kinases with LRR and bulb lectin domains. This provides strong evidence that evolution has worked in the crown group of eukaryotes to produce rather similar solutions for multicellularity on independent occasions.

Conclusion

These consistent results for C. elegans and A. thaliana allow the reconstruction of the possible evolutionary events that lead to multicellularity. The majority of the regulatory domains were present in the common ancestor of the crown group but the general lack of conservation of domain architectures suggests that they were independently recruited and expanded along the different lineages. This is particularly striking in the case of the transcription factors where there have been additional specific domains in each lineage. Further, it is possible that specific transcriptional regulation developed under rather distinct selective forces in different eukaryotic lineages: in yeast for regulation of temporal states and in animals and plants for spatial and developmental diversification of cell states.

In reconstructing the origin of multicellularity, one can imagine a ground state in the form of a colony of identical cells. Certain changes that restricted the expression of ancestral transcription factors to distinct subsets of cells in this colony could have triggered off spatial differentiation. Selection of organisms with such systems could have led to the fixation of interacting protein networks and led to the origin of forms with an increased number of cell types.
Such a process would have provided a selective opportunity for the expansion of transcription factors (for differentiation), signaling molecules (to respond to intercellular communication) and adhesion molecules (to hold the cells together). Thus, we see these molecules as prominent components of the proteomes of complex multicellular eukaryotes such as C. elegans and A. thaliana. Although the issue of primary derivation of these novel regulatory functions like adhesion and intercellular signaling could be raised, as mentioned above, recruitment of pre-existing domains appears to have played the a major part. Those domains that are entirely new are largely based on stabilizing disulfide bonds or chelated metal ions, suggesting that they could have arisen relatively easily by de novo. Thus, complex as multicellular eukaryotes may appear, a set of rather simple steps could have lead to their origins.

The Drosophila genome is expected to be fully sequenced by the end of this year, with those of Homo sapiens and Distylium docordum to follow. Analysis of these genomes are likely to provide exciting insights regarding the relative evolutionary time in which the regulators of multicellular development and patterning emerged.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:
** of special interest
** of outstanding interest

1. ** Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M:**
   
   The approach to obtaining the complete sequence and associated caveats.

2. ** Dickinson JR, Schweizer ME:**
   
   A recent book providing details regarding the physiological states of the cells, embryos, and evolution: toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability. Blackwell Science 1992

3. ** Gerhart J, Kirschner MW:**
   

4. ** 21**

5. ** 20**

6. ** 19**

7. ** 18**

8. ** 17**

9. ** 16**

10. ** 15**

11. ** 14**

12. ** 13**

13. ** 12**

14. ** 11**

15. ** 10**

16. ** 9**

17. ** 8**

18. ** 7**

19. ** 6**

20. ** 5**

21. ** 4**

22. ** 3**

23. ** 2**

24. ** 1**

This paper describes the provenance of the ‘eukaryotic’ signaling domains. The intracellular occurrence of von Willebrand factor A domains in the TPR p44 subunit is reported for the first time and enumeration of these domains in complete genomes is provided.

4. **Penning CP, Aravind L, Schultz J, Boek P, Koornneef E:**
   

5. ** 285**

6. ** 282**

7. ** 278**

8. ** 272**

9. ** 265**

10. ** 259**

11. ** 253**

12. ** 247**

13. ** 239**

14. ** 233**

15. ** 227**

16. ** 221**

17. ** 215**

18. ** 209**

19. ** 203**

20. ** 197**

21. ** 191**

22. ** 185**

23. ** 179**

24. ** 173**

25. ** 167**

26. ** 161**

27. ** 155**

28. ** 149**

29. ** 143**

30. ** 137**

31. ** 131**

32. ** 125**

33. ** 119**

34. ** 113**

35. ** 107**

36. ** 101**

37. ** 95**

38. ** 89**

39. ** 83**

40. ** 77**

41. ** 71**

42. ** 65**

43. ** 59**

44. ** 53**

45. ** 47**

46. ** 41**

47. ** 35**

48. ** 29**

49. ** 23**

50. ** 17**

51. ** 11**

52. ** 5**

53. ** 1**

** Origin of multicellular eukaryotes — insights from proteome comparisons Aravind and Subramanian 693

HMMER program that can be downloaded from the World Wide Web at:

http://hmmer.wustl.edu/

Penning CP, Schultz J, Boek P, Koornneef E:**


Excellent system for the annotation of signaling domains in proteins based on HMMs derived from high-quality alignments. Can be accessed on the World Wide Web at:

http://smart.embl-heidelberg.de/SMART/

Cheruit SA, Aravind L, Shoemaker GC, Bal CA, Koornneef E, Dwight SS, Harman MA, Dolinski K, Minton S, Smith T et al:**


The first comparison of the complete proteomes of two eukaryotes. The authors describe the lineage-specific domain expansions, conservation of the core functions and the use of PSI-BLAST sequence profiles in establishing counts for proteins containing particular domains.

Sluder AE, Mathews SW, Hough D, Yin VP, Mana CV:**


Divergence time estimates and turnover of the domain families in the nematode Orygnemio arenarium. The study demonstrates the utility of the mammalian-specific expansion that has occurred within this superfamily.

Pappasouris VE, Silver EM:**


Describes the diversity of the T-Box transcription factors in the animals.

Hobert O, Ruvkun G:**


Describes the interesting case of the alternative domain organizations of two HTH domains comprising the Paxed Box. The carboxy- and amino-terminal HTH domains have been shown to occur independently of each other as well as in the form of fusions as in the classic Paled domains.

Ruvkun G, Hobert O:**


Provides a detailed summary of the conservation of developmental pathways in the nematodes. Some of the conclusions reported here need to be reconsidered due to use of an outdated sequence set.

Aravind L, Koornneef E:**


This paper describes the provenance of the ‘eukaryotic’ signaling domains. The intracellular occurrence of von Willebrand factor A domains in the TPR p44 subunit is reported for the first time and enumeration of these domains in complete genomes is provided.

Penning CP, Aravind L, Schultz J, Boek P, Koornneef E:**


Describes the intracellular occurrence of von Willebrand factor A domains in the TPR p44 subunit.

Baikman A, Chothia C:**


Describes the interesting case of the alternative domain organizations of two HTH domains comprising the Paxed Box. The carboxy- and amino-terminal HTH domains have been shown to occur independently of each other as well as in the form of fusions as in the classic Paled domains.

Shoelson SE:**


Describes the interesting case of the alternative domain organizations of two HTH domains comprising the Paxed Box. The carboxy- and amino-terminal HTH domains have been shown to occur independently of each other as well as in the form of fusions as in the classic Paled domains.

Hurley JH:**


Provides an introduction to the recent developments regarding the cyclic nucleotides generating enzymes.

Ranganathan R, Ross EM:**


Reviews the role of POZ domains in acting as the sub-membrane scaffolds in animal signaling systems.

Suva M:**


Describes the interesting case of the alternative domain organizations of two HTH domains comprising the Paxed Box. The carboxy- and amino-terminal HTH domains have been shown to occur independently of each other as well as in the form of fusions as in the classic Paled domains.

Paton EE, Wellner AR, Tynan M:**


Provides a good introduction to the significance of the F-Box in ubiquitin-dependent signaling. The role of the F-box module in targeting the ubiquitin-conjugating system is described.


