CHAPTER 13

Computational reconstruction of ancestral genomic regions from evolutionarily conserved gene clusters

Etienne G.J. Danchin, Eric A. Gaucher, and Pierre Pontarotti

13.1 Introduction

Reconstruction of ancestral genomic features can be considered on multiple evolutionary scopes and at different levels of biological sequence information. For instance, one could anticipate the reconstruction of genomic features for the last common ancestor of all species on Earth, last universal common ancestor or LUCA, whereas others would focus on reconstructing these features in the last common ancestor of vertebrates and/or arthropods. In an analogous manner, biological sequences themselves can be divided into subcategories as a function of their nature or their scale. It is possible to consider reconstructing ancestral genes, ancestral proteins, ancestral retro-elements, ancestral chromosomes, or even an ancestral genome. We present here our conceptual and computational approach for reconstructing gene clusters, with a particular emphasis on the major histocompatibility complex (MHC) region. We anticipate that our approach will be extended, and coincide with technological advancements allowing reconstructionists to synthesize ancient genomes in the laboratory.

13.2 Small-scale reconstructions

On the smaller scale, representing individual sequences (i.e. gene, protein, mobile element, etc.), reconstruction of ancestral biological sequences can go beyond the conceptual level and lead to a physical reconstruction of the deduced ancestral sequence. Indeed, several research articles relate physical reconstruction of biological sequences based on phylogenetic reconstructions to ancient organismal behaviors, as reviewed in various chapters in this book.

13.3 Larger-scale reconstructions

Alternatively, larger-scale biological sequence reconstructions are concerned with ancient chromosomes, genomic regions, and genomes. Fewer studies, however, have been presented on this scale (Blanchette *et al*., 2004). Moreover, they do not go beyond the conceptual level *in silico* because (for the moment) technology does not allow extension towards physical reconstructions. A logical step towards realizing an ancestral genome consists first of inferring the gene content of the ancestral organism.

13.3.1 Ancestral gene content reconstruction

Several authors have recently evaluated the number of genes or proteins most likely present in the ancestors of different animal phyla. Koonin *et al*. (2004) performed an in-depth comparative analysis of whole proteomes from seven different eukaryotic species. Based on identified clusters, and on a study of the evolution of these species, they inferred the gene set that was probably present in

the last common ancestor of the eukaryotes to consist of at least 3413 gene families. In a similar manner, they also evaluated the gene set for each internal node of the phylogeny of these seven species and, for example, they estimated that the last common ancestor of all bilaterian species had at least 5313 gene families. Using a similar approach, Hughes and Friedman (2004) compared complete proteomes of various bilaterian species (insects, vertebrates, and nematodes), and estimated that approximately 2100 protein families were present in the last common ancestor of these taxa (*Urbilateria*).

It is interesting to note here that these two analyses provide very different estimates (more than 2-fold) of the ancestral bilaterian proteome size. This difference can be explained by the fact that the set of species used to define the size of the ancestral proteome was not the same for the two analyses. Moreover, the definition of gene families between the two analyses was slightly different, and also the methods used to deduce ancestral gene content from clusters of conserved genes were not identical.

Both these approaches evaluated clusters of putative orthologous groups of protein families by all-against-all pairwise comparisons of proteins between the different species, but did not systematically test the orthology relationships between these genes by phylogenetic analysis. Sequence similarity-based approaches can misguide in some instances where evolutionary relationships between genes are particularly complex whereas phylogenetic analysis tends to resolve such complex cases (Danchin, 2004; Jordan *et al*., 2004; Gouret *et al*., 2005). Nevertheless, as explained by the authors, phylogenetic analysis for genome-wide comparisons can also be erroneous and remains labor-intensive. Even if these two analyses are likely to include false positive and negatives, they represent the most reliable estimations of ancestral gene and protein sets to date.

These studies evaluate the putative gene or protein content in the ancestor of various phyla, at the largest scale possible, through comparative analysis. Although similar analyses have been performed for Bacteria (Kunin and Ouzounis, 2003), we focus here on ancestral eukaryotic genome content.

13.3.2 Reconstruction of ancestral genomic organization

Several methods and analyses have been developed to reconstruct ancestral genome organization. For example, Bourque and Pevzner (2002) developed a method to decipher ancestral gene orders based on the comparison of gene order between modern species. These authors then presented a follow-up reconstruction of the genomic organization of the rodent ancestor from mouse and rat based on comparison of conserved genomic blocks and their relative order (Bourque *et al*., 2004). This genomic reconstruction included both coding and non-coding chromosomal regions but did not consider genomic regions that had been duplicated. Nor did it give information about the organization of genes inside the genomic blocks. More recently, Bourque *et al*. (2005) expanded their original method and proposed a reconstruction of the ancestral genome organization of the murid rodent ancestor, and of the mammalian ancestor. This latest analysis provides an opportunity to reconstruct gene content and organization inside the ancestral genomic blocks by considering comparisons at the coding regions level. In parallel, and using a similar approach, Jaillon *et al*. (2004) proposed a reconstruction of the ancestral karyotype of the vertebrates through comparison between the teleost fish *Tetraodon nigroviridis* and the human genome.

These analyses predicted a putative genomic organization in mammal, rodent, and vertebrate ancestors at the whole-genome scale. However, both of the analyses used reciprocal best-BLAST (Altschul *et al*., 1997) hit approaches to decipher orthology relationships (known to be problematic) and neither study considered duplicated regions and genes. Due to the limited number of whole genomes available for comparison, these analyses certainly missed genes or regions that were lost multiple times in different lineages, and thus ancestral reconstructions lacked these elements. We surmise that increasing the number of genome comparisons will lead to greater resolution.

13.3.3 Reconstruction of ancestral genomic regions through comparisons of evolutionarily conserved gene clusters

The reconstruction of ancestral biological features achieved in our research group to date is at an intermediate scale between individual sequences (genes, proteins, mobile elements, etc.) and large-scale reconstruction (whole ancestral karyotypes, genomes, or proteomes). We proposed the reconstruction of genomic regions at the level of their ancestral gene content (Danchin *et al*., 2003; Danchin, 2004; Danchin and Pontarotti, 2004a, 2004b) through the comparison of evolutionarily conserved gene clusters. Thus far, our conceptual reconstructions have not included predictions on the organization of genes (i.e. order and orientation) inside the ancestral regions, but are rather predictions of ancestrally grouped genes irrespective of their relative organization inside the clusters.

Our initial analyses focused on reconstructing regions in the last common ancestor of the euchordates (Danchin and Pontarotti, 2004b; named *Ureuchordata*) and in the last common ancestor of the bilaterians (Danchin *et al*., 2003; Danchin, 2004; Danchin and Pontarotti, 2004a, 2004b; named *Urbilateria*). The most obvious way to expand these initial analyses of ancestral genomic information content is to compare the genomic *organization* of conserved regions that are suspected to have originated from a common ancestral region.

Reconstruction of ancestral genomic clusters as far back as the last common ancestor of all bilaterian species (*Urbilateria*) has been possible through the comparison of genomic regions whose gene composition was evolutionarily conserved between Protostomes (like *Drosophila melanogaster*) and Deuterostomes (like *Homo sapiens*). Evolutionarily conserved genomic regions were identified between Protostomes and Deuterostomes prior to reconstructing putative ancestral clusters. We first started from selected regions in the human genome for which we had evidence of evolutionary conservation in vertebrates. These selected regions of the human genome consisted of relatively well-conserved paralogous gene clusters that had been shown previously to originate from a common ancestral region after duplication (Abi-Rached *et al*., 2002; Vienne *et al*., 2003a). From these clusters, we next retrieved genes that appeared to constitute signatures of evolutionary conservation. These so-called signature genes had to fulfill several criteria, in that they must be present in at least one copy in one of the paralogous regions and the estimation of their duplication date should be in a consistent time window. Orthologs to these anchor genes were then searched for in the genomes of protostomian species (i.e. *Anopheles gambiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans*) by a systematic phylogenetic analysis. We retrieved genomic locations of each protostomian gene having a human ortholog. For each protostomian genomic segment containing at least two orthologs and spanning less than 2 Mb, a statistical test was applied. The appropriate statistical test allows us to distinguish significant conservation from conservation by chance.

13.4 Choice of candidate regions

Our previous analyses of bilaterian ancestral genomic reconstructions relied on ancient duplicated clusters that today have remained structurally conserved. These clusters resulted from two rounds of duplication from a unique ancestral region after the divergence between cephalochordates (amphioxus, *Branchiostoma floridae*) and craniates (hagfishes plus vertebrates), and before the emergence of gnathostomata (jawed vertebrates). These paralogous regions retained significant conservation of gene content despite hundreds of millions of years of divergence from their common ancestral state.

The two sets of quadruplicated regions studied were the MHC and its paralogous regions, and the 8–10–4-5 regions. For both sets, data suggested the existence of an ancestral region (at least early in chordate history) from which they originate, and derived after *en bloc* duplications. Indeed, conservation of gene clustering can still be observed between the paralogous regions inside a given quadruplicated set (Abi-Rached *et al*., 2002; Vienne *et al*., 2003a). As a consequence, the two sets of four paralogous regions we observe today in vertebrate genomes may represent echoes of a conserved

common ancestral cluster. In our objective towards reconstructing ancestral regions, our preliminary observations placed these quadruplicated regions as obvious candidates to look for further conservation in other species within the tree of life.

We hypothesized that these two sets of quadruplicated regions in vertebrates (Deuterostomes) may have diverged from a more ancient genomic cluster, possibly as distant as Protostomes. The remainder of this chapter will focus on the MHC and its three paralogous regions, since the strategy and approach used for the 8-10-4-5 regions are analogous.

13.4.1 The MHC and its paralogous regions

The MHC region is located in the human genome on chromosome 6p21.3. This genomic region of approximately 2 Mb contains genes that are involved in the immune response. For instance, PSMB8 and PSMB9 encode two subunits of the immunoproteasome (a multimeric complex which cleaves peptides to a specific size for presentation at the cell surface), and C4 encodes a subunit of the complement system (a 30-protein system involved in immunological response, anaphylaxis, and cell destruction). Other genes with no clear reported role in immunity are also present in this region. For example, retenoid X receptor (RXR) B is a co-activator that increases the DNA-binding activity of retinoic acid receptors (RARs) whereas PBX2 encodes a protein with a homeobox domain but whose function is not well documented.

Three other regions of the human genome (chromosomes 1p22–p11, 9q33–q34, and 19p13) contain clustered copies (paralogs) of some of the genes present in the MHC region. This observation was initially made by Kasahara *et al*. (1996, 1997), who defined three MHC-like regions in the human genome in addition to the original MHC region on chromosome 6p21.3. These three MHC-like regions have been predicted by Abi-Rached *et al*. (2002) to have been the result of two rounds of *en bloc* duplication from an ancestral region. A schematic representation of the MHC region as well as its three paralogous conserved regions is presented in Figure 13.1. These four paralogous clusters arose through duplication from their common ancestral region around 700 million years ago (Abi-Rached *et al*., 2002). During millions of years of evolution these regions may have undergone fixation of several rearrangements. Among these rearrangements, gene loss and translocations could be invoked to explain why not all members of quadruplicated genes are still present as four copies in the quadruplicated regions. For example, in the RXR family, one paralogous copy is found on each of chromosomes 6, 1, and 9 (respectively RXRB, RXRG, and RXRA) but no paralogous copy is present within the fourth region (on chromosome 19). The same type of loss pattern is also found for other genes not listed here. In some cases, losses can be more extended and leave only two remaining copies (as for AGPAT family; 1-acylglycerol-3-phosphate O-acyltransferases 1 and 2). Note that at this stage it is difficult to state whether singleton genes are the remains of quadruplicated genes that experienced multiple losses, or whether they represent a single-copy gene translocated into these regions after the *en bloc* duplications and subsequent divergence from the common ancestral region. An important point that must be specified is that the relative order of genes along the four regions of paralogy is not conserved between the MHC and any of its three paralogous regions. Thus, the only feature that characterizes these regions is a common clustering of paralogous genes regardless of their relative order.

13.5 Conservation in other species

Anchor genes representing signatures from the two sets of vertebrate quadruplicated regions (as defined above) were used to identify potentially conserved clusters in other species. The species that have been tested for conservation were chosen according to the following criteria: their genomes are completely sequenced, assembled, and annotated to allow retrieval of gene locations along the genome. The selected species were *Drosophila melanogaster*, *Anopheles gambiae* (two dipteran insects), and *Caenorhabditis elegans* (a nematode). These three species are all bilaterian species belonging to the protostomian group. Moreover, while still debated today for nematodes (Blair *et al*., 2002; Copley *et al*., 2004; Telford, 2004a, 2004b;

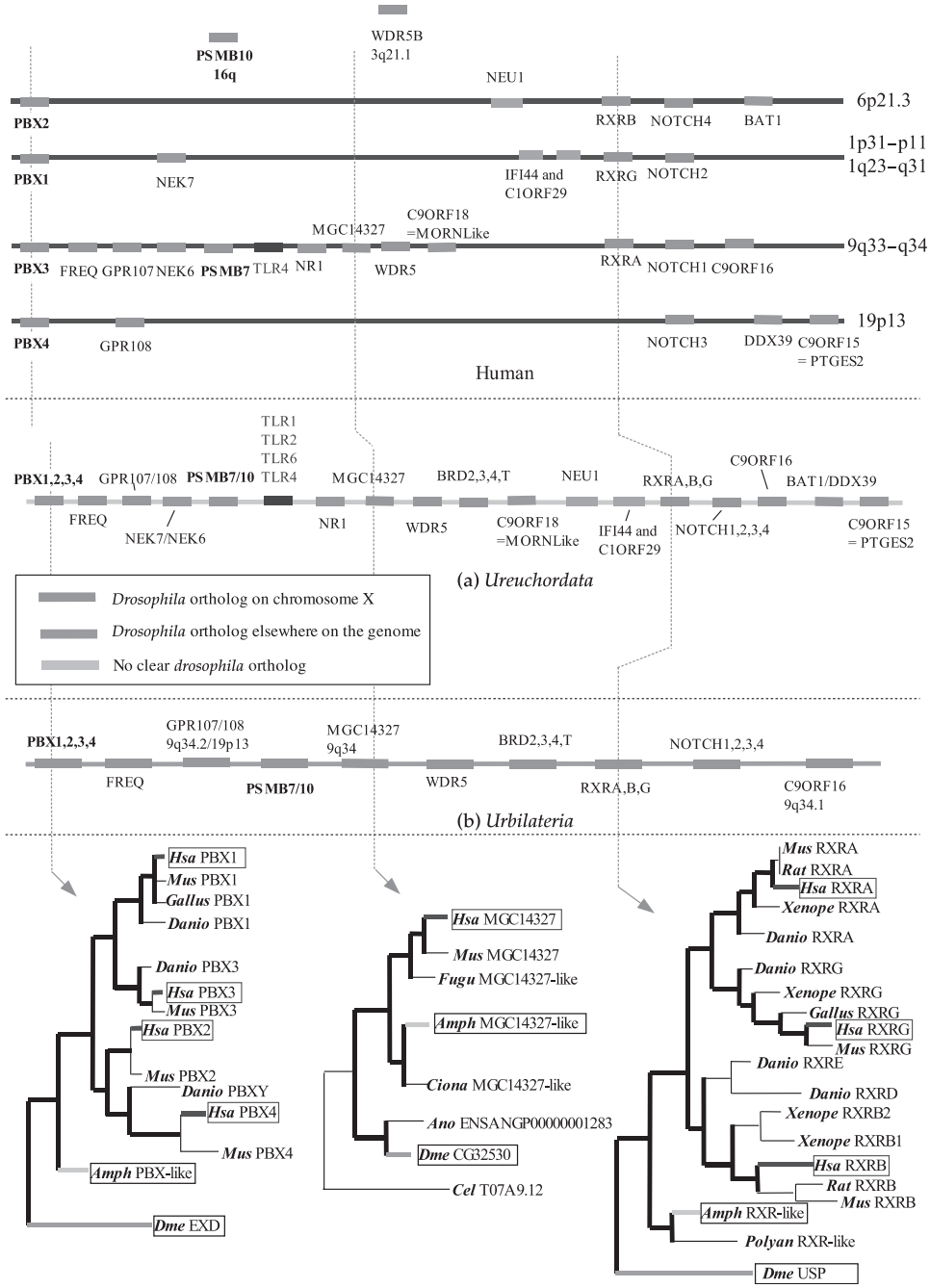

Figure 13.1 *Ureuchordata* and *Urbilateria* proto-MHC reconstructions. Top panel: distribution of the 18 conserved gene families between human and amphioxus on the human MHC and paralogous regions. (a) Minimal reconstruction of the putative ancestral region in *Ureuchordata*. (b) Reconstruction of a minimal region in *Urbilateria* based on conserved clustering in *Drosophila*. Bottom panel: three examples of phylogenetic trees for three gene families presenting different patterns of gene presence or absence. Note: the actual organization (i.e. order and orientation) of genes on the reconstructed ancestral regions is not known and probably rearranged; we chose to represent homologous genes in the same order on the various different regions so that their homology relationships are easier to read. *Hsa*, *H. sapiens*; *Amph*, amphioxus (*Branchiostoma floridae*); *Dme*, *D. melanogaster*; *Ano*, *Anopheles gambiae*; *Cel*, *C. elegans*; *Xenope*, *Xenopus laevis*; *Polyan*, *Polyandrocarpa misakiensis*; *Mus*, *Mus musculus*; *Danio*, *Danio rerio*; *Fugu*, *Takifugu rubripes*; *Gallus*, *Gallus gallus*; *Rat*, *Rattus norvegicus*; *Ciona*, *Ciona intestinalis*.

Wolf *et al*., 2004) they may all be in the same group of protostomians, the ecdysozoans. Additionally, we also used partial genomic information available for other species such as one cephalochordate (*Branchiostoma floridae*) and two urochordates (*Ciona intestinalis* and *Ciona savignyi*) in the case of the MHC and its paralogous regions.

The MHC region and its three paralogous regions in human are widely conserved in other primates (chimpanzee), mammals (mouse, cattle, and rat), and vertebrates (fugu and zebrafish; Flajnik and Kasahara, 2001). Outside the vertebrate lineage, conservation has been shown with parts of the cephalochordate amphioxus genome (Abi-Rached *et al*., 2002).

In addition, conservation has been reported (Trachtulec *et al*., 1997; Trachtulec and Forejt, 2001) with *D. melanogaster* chromosome X, *C. elegans* chromosome III, and parts of snake and *Schizosaccharomyces pombe* genomes. These latest instances of conservation were not confirmed by a phylogenetic analysis assessing orthologous relationships between genes of the various species considered. Moreover, the statistical test performed for these analyses did not consider heterogeneity of gene distribution along the genomes. Therefore, these examples needed to be confirmed and completed with new genomic data available. We present here a summary of our analyses (Danchin *et al*., 2003; Danchin and Pontarotti, 2004b). We confirm some of the previously reported examples of conservation, reject others, and expand the knowledge of conservation of gene organization inside these regions.

13.5.1 Conservation in euchordates

Previous work on MHC and its three paralogous clusters have suggested conservation of the cluster between jawed vertebrates and amphioxus (*B. floridae*; Flajnik and Kasahara, 2001; Abi-Rached *et al*., 2002; Vienne *et al*., 2003b). In combination with follow-up studies (Castro *et al*., 2004), we were able to identify conserved clusters of 18 families of orthologous genes between human and amphioxus. Moreover, the conservation of this clustering is statistically significant (Abi-Rached *et al*., 2002; Vienne *et al*., 2003b; Danchin and Pontarotti, 2004b). Altogether, these results demonstrate the existence of an MHC-like region in the amphioxus.

Furthermore, as the MHC and its three regions of paralogy are widely conserved among vertebrates, we can hypothesize that the conservation between vertebrates and cephalochordates reflects the existence of a proto-MHC region before the divergence between these lineages. This hypothesis is consistent with the fact that the four regions of paralogy found in vertebrates duplicated *en bloc* from a common ancestral region after separation between the cephalochordate (amphioxus) and craniate (hagfishes plus vertebrates) lineages, but before the emergence of the gnathostomata (jawed vertebrates; Abi-Rached *et al*., 2002).

Conserved gene clustering with the MHC and its three paralogous regions can thus be deciphered throughout the euchordate lineage (cephalochordates and craniates) and such conservation suggests that a proto-MHC cluster probably existed in the last common ancestor of these two lineages. We name the last common ancestor of all euchordates *Ureuchordata* (Danchin and Pontarotti, 2004b) in reference to *Urbilateria*, which is the last common ancestor of all bilaterian species. A reconstruction of this *Ureuchordata* proto-MHC region is considered further in the next sections, and in Danchin and Pontarotti (2004b), by comparing the genomic organization of cephalochordate and vertebrate MHC-like regions.

13.5.2 Conservation in chordates

Conservation in the euchordate lineage is clear, as shown in the previous section. Based on this observation, we investigated whether conservation could be revealed more widely in the chordate lineage. We thus compared genomic information from euchordate species with partial genomic information available for two urochordate species (*Ciona intestinalis* and *Ciona savignyi*), detailed in Danchin and Pontarotti (2004b). As the last common ancestor of species from the euchordate and urochordate lineages is the last common ancestor of all chordates, information on genomic organization from *Ciona* species would provide additional data for deciphering the ancestral genomic organization

in *Ureuchordata*. Moreover, this could provide evidence for a more ancestral pre-existing region in the last common ancestor of all chordates.

The genomes of the two *Ciona* species are both fully sequenced and assembled into scaffolds of various lengths whose relative positions, however, are unknown to date. If significant conservation can be found in the genome of these urochordates, we do not expect it to extend further than scattered pieces of small conserved clusters. We nevertheless identified traces of conservation for the MHC-like regions between the amphioxus, human, and either one or both the two *Ciona* species. Indeed, several sets of *Ciona* orthologs are grouped on the same *Ciona* scaffolds in a manner similar to genes clustered in euchordate MHC-like regions (in human or amphioxus; Danchin and Pontarotti, 2004b). These sets consist of two scaffolds of three co-localized genes in *Ciona savignyi* in conjunction with one scaffold of four co-localized genes, one scaffold of three co-localized genes, and two scaffolds of two co-localized genes in *Ciona intestinalis* (Danchin and Pontarotti, 2004b).

Unfortunately, as none of the two Ciona genomes are sufficiently assembled, we could not statistically test the significance of this conservation between euchordates and urochordates. However, as *Ciona* genomic information becomes more advanced, we should be able to test whether this conservation is significant and may reveal the existence of a proto-MHC region at the base of chordate evolutionary history.

13.5.3 Conservation in bilateria

Conservation of an MHC-like region is clear and statistically significant inside the euchordate lineage between cephalochordates and vertebrates (Danchin and Pontarotti, 2004b). This conservation suggests the pre-existence of a proto-MHC region in *Ureuchordata*, the last common ancestor of all euchordates. In parallel, we revealed traces of conservation between urochordates, vertebrates, and cephalochordates (Danchin and Pontarotti, 2004b). The significance of this conservation could not be evaluated but may also indicate conservation of an MHC-like region in urochordates. We needed to identify conservation in Protostomes, however, to attain our goal of reconstructing ancestral genomic organization back to the origin of the Bilateria. As far as this is concerned, we showed statistically significant conservation of MHC and paralogous regions clustering in *Drosophila melanogaster* with both vertebrates (human) and cephalochordates (amphioxus). Moreover, we also showed statistically significant conservation of the clustering between *Drosophila melanogaster* and *Anopheles gambiae* (Danchin *et al*. 2003; Danchin and Pontarotti, 2004b). Altogether these results show conservation of an MHC-like genomic region organization between Deuterostomes and Protostomes. The last common ancestor of all these species is *Urbilateria*. Thus, the observed conserved gene clusters may represent orthologous regions that originated by speciation from a common ancestral region in *Urbilateria*.

13.6 Significance and hypotheses concerning conservations

We identified conservation of genes clustering in several species including Deuterostomes and Protostomes for the MHC and its three paralogous regions (and for the 8–10-4–5 quadruplicated regions (Danchin and Pontarotti, 2004a)). Deuterostomes and Protostomes separated more than 700 million years ago from the last common ancestor of bilaterian species (Douzery *et al*., 2004). We can address the question of significance and the expectation of observing conservation despite such evolutionary divergence between the species considered here.

Several hypotheses can be considered to explain conservation between such phylogenetically distant species. The first hypothesis, even if unlikely, is that the conserved genomic organization is due to chance and is not biologically significant. As described previously, the significances of the conserved clusters we deciphered were all evaluated by a statistical test. In all the cases we tested the following null hypothesis (H_0): the distribution of species B orthologs to species A genes (present in a given region X) is random along the genome of species B and does not reflect significant conservation. As further detailed in Danchin *et al*. (2003) and Danchin and Pontarotti (2004a, 2004b) the statistical test allowed the rejection of the null hypothesis in all cases except for comparisons with

C. elegans. These tests thus suggest that all the conservations (except for the nematode) were biologically significant (Danchin *et al.*, 2003; Danchin and Pontarotti, 2004a, 2004b). The null hypothesis of similarity in gene content by chance could be rejected with high significance, and this is particularly unexpected for clusters between chordates and dipteran since they diverged more than 700 million years ago. Genomes can have accumulated numerous rearrangements in the different species since their divergence from their last common ancestor. The more ancient the divergence is, the more likely these genomes are to be differentially organized.

Two alternative hypotheses can explain conservation when the null hypothesis has been rejected. This can either be the result of evolutionary conservation from an ancestral cluster, or be due to convergent evolution with positive selection driving similar genome organization/content.

13.6.1 Convergence with positive selection

An apparent role of shared ancestry between Deuterostomes and Protostomes may be the result of convergence with positive selection. Here, the genes considered in the two sets of conserved regions may not be ancestrally clustered, but rather, the genes grouped together within chordate (for Deuterostomes) and dipteran insect (for Protostomes) lineages separately.

Under this hypothesis, reconstruction of ancestral genomic regions should not be considered, as the conserved clusters we observe do not represent traces of the existence of clusters in the ancestor of the considered species. It is interesting, however, to investigate the evolutionary forces that could have favored these genes to independently cluster in two different lineages. We can imagine that particular positive selection acting on these genes favored their clustering into limited regions. Such a hypothesis could be tested if sufficient expression and functional data for these genes are available. Few functional or expression data for these genes in the different species considered here are currently available, and it is difficult to test this hypothesis at the moment. The logic of this argument is that convergence of location would likely be driven by co-expression if there were a positive selection pressure driving it.

13.6.2 Likelihood of the hypotheses

In our goal towards reconstructing ancestral genomic clusters in *Urbilateria*, it is necessary to consider the likelihood of the different hypotheses. Indeed, the reconstruction analysis is only possible under the hypothesis that the conserved clusters derived from a common ancestral region. As shown above, the hypothesis of similarity by chance can be rejected and thus the two alternative hypotheses that remain are the hypotheses of ancestral conservation and convergence with positive selection.

13.7 Reconstruction of ancestral regions

Based upon the hypothesis that the conservations between Protostomes and Deuterostomes that we observe constitute traces of inheritance from a common ancestral region, we propose conceptual reconstruction of the putative ancestral region from which they are derived. The general strategy for these reconstructions consists of inferring the presence of a given gene in the ancestral region based on its presence in both the corresponding conserved regions of the compared species. Using this approach, we necessarily provide a minimal reconstruction which only includes genes that are both still present in the two compared regions. As a consequence, genes ancestrally present in the region but that were lost in one or both of the compared conserved regions will not be included in these reconstructions. Similarly, genes initially present in the ancestral region but that were translocated to new locations after speciation are also absent from the reconstruction. Future comparisons to other phyla will help determine the ancestral presence or absence of uncertain genes discussed above. In all, the reconstructions we propose should be viewed as a minimal set of genes whose clustering is conserved throughout evolution and which are thus probably a remnant of ancestral gene clusters.

Reconstructions at different evolutionary scales can be considered, and we propose reconstruction

of an ancestral MHC region, both in the ancestor of all euchordates and in the ancestor of all bilaterian species. For the MHC and its three paralogous regions we benefit from genomic-organization data in a wide variety of vertebrate species and from additional information concerning the amphioxus, as well as partial data on urochordates. Based on these data we can propose a reconstruction of the proto-MHC in the ancestor of all euchordates. In addition, comparisons with conserved cluster data from insects will generate inferences for *Urbilateria*.

13.7.1 Euchordates

We identified 18 families of orthologous genes whose clustering is conserved between vertebrates and cephalochordates (Danchin and Pontarotti, 2004b). We propose that these conserved clusters echo an ancestral cluster in which 18 ancestral genes were already grouped in the last common ancestor of all euchordates, namely *Ureuchordata*. As the duplications that gave rise to the gene families in vertebrates occurred after the separation between cephalochordates and gnathostomes (Abi-Rached *et al*., 2002; Vienne *et al*., 2003a; Danchin and Pontarotti, 2004b), we can deduce that genes were single-copy in the ancestor of these two species. A minimal reconstruction of the ancestral cluster in *Ureuchordata* is presented in Figure 13.1a.

13.7.2 *Urbilateria*

Comparisons of conserved gene clusters between vertebrates and insects allowed us to propose a putative proto-MHC cluster of 19 genes in the last common ancestor of bilaterians (Danchin *et al*., 2003). Follow-up analyses based on additional data for conservation of an MHC-like region throughout euchordates proceeded with a comparison of the putative ancestral *Ureuchordata* reconstructed proto-MHC region to the *Drosophila* genome. Based on this analysis we identified 10 families of orthologous genes whose clustering is conserved between vertebrates, cephalochordates, and insects. From these 10 gene families we deduced a core ancestral cluster of 10 genes that was probably present in *Urbilateria*. These 10 genes remain clustered in modern species despite approximately 700 million years of evolution for each derived cluster (Danchin and Pontarotti, 2004b). A representation of this ancestral cluster is illustrated in Figure 13.1b.

For genes present in the ancestral region reconstructed by Danchin *et al*. (2003) but not on the reconstructed region of Danchin and Pontarotti (2004b), these constitute good candidates to check for their presence in the cephalochordate MHC-like chromosomal region. Unfortunately no supplemental genomic data are available today for the amphioxus, but when such data are released we can consider testing this hypothesis. If this is demonstrated, such genes can be reintroduced in both the reconstructed ancestral *Ureuchordata* and *Urbilateria* proto-MHC regions, leading to a more accurate and complete reconstruction.

13.8 Discussion and perspectives

For the two sets of quadruplicated regions analyzed (MHC and 8-10-4-5), we developed a method combining phylogenetic analysis and statistical testing that allowed identification of a set of statistically significant conserved gene clusters between phylogenetically distant species. Based on these conservations, we then proposed reconstruction of the minimal gene content of the corresponding region in the last common ancestor of the compared species, *Urbilateria*. In order to make additional progress in the reconstruction of the genome of our distant Bilaterian ancestor, several points can be considered. The first one concerns improvement of the reconstruction methods and the development of an algorithm to evaluate the likelihood of presence/absence of a given gene in an ancestral region. An additional point undoubtedly consists of enriching the analysis by including new genomic information from phylogenetically informative species to improve the reliability and sensitivity of reconstructions. Lastly, we can also consider automation of the process with inclusion of the improved reconstruction method, the likelihood algorithm, and new species data. Such automation would allow high-throughput treatment of potential regions of evolutionary conservation, and thus provide advanced tools for the reconstruction of the genome of our distant bilaterian ancestor.

13.8.1 Improvements to the reconstruction method

To date, reconstructions of ancestral regions that we have proposed have been based on the manual examination of genes present in all the evolutionarily conserved clusters. The basic concept is that the minimal common set of genes present in all the regions of conserved syntenies we compare were ancestrally present in the regions from which it originated. Whereas the co-presence of genes in multiple regions of conserved synteny provides strong support for their ancestral presence, there are several drawbacks. The first and most important one is that this method does not provide any statistical value (or score) for the likelihood of ancestral presence or absence of a given gene in the deduced corresponding ancestral region. The second is that by using such an approach we necessarily miss genes that were translocated to a new position or lost. A statement about the ancestral presence of a gene in an ancestral region may require comparisons to a third species. Such cases will also require a weighting scheme for the presence of homologs in additional taxa. A likelihood value for the ancestral presence of the gene in the ancestral corresponding region should then be assigned. This likelihood value should vary as a function of the phylogenetic pattern of presence/absence of the gene in an evolutionarily conserved cluster in different species. Lastly, some cases require that we consider the potential of convergent translocation of genes into two regions of conserved synteny between two species while they were actually not present in the corresponding ancestral region. Once again a likelihood value should be assigned with the use of an appropriate statistical test in conjunction with comparisons of additional species. Several methods, using parsimony or likelihood have been developed to evaluate the probability of presence or absence of a given gene in an ancestral genome with regards to their phylogenetic patterns (Kunin and Ouzounis, 2003; Koonin *et al*., 2004). Such methods could be adopted to reconstruct ancestral gene content by adding a term for the conservation of genomic location.

13.8.2 Automation of the pipeline

We previously developed an automated computational platform, termed FIGENIX (Gouret *et al*., 2005), dedicated to biological sequence analysis. In a collaborative effort with Philippe Gouret and Virginie Lopez Rascol, we are developing a multiagent system, named CASSIOPE, to find all statistically significant conserved clusters in other species and inside the query species (for possible regions of paralogy), test the convergence and inheritance hypotheses, and then propose a putative reconstruction of the corresponding ancestral region at various nodes of the bilaterian tree of life. The whole process will be based upon the following main steps:

- All the genes present in a queried genomic region are extracted, and homologs are automatically searched for in all the other species whose genomes are fully sequenced (orthologs) and reside within the genome of the search species (paralogs). At this step, the FIGENIX platform will be used to automatically detect homologs based upon robust phylogenetic reconstruction.
- Genomic locations of all homologs found in other species and in the search species are extracted. Selection of all genomic segments, defined by at least two homologs on the same DNA molecule (i.e. same chromosome) is performed, including a test of statistical significance with two maximizations (max significance or max cluster length).
- Ancestral clusters are reconstructed using parsimony according to the phylogenetic pattern of presence or absence of genes in the conserved clusters among different phyla.

With such automation of the process, it will be possible to progress faster toward reconstructing the genome of *Urbilateria*. Ultimately, ancestral reconstructions at higher levels of organization will be considered until the goal of whole-genome reconstruction is realized.

13.8.3 Beyond genomes

Additional ancestral features can be considered for reconstruction beyond the reconstruction of

ancestral genomic clusters. By comparing the proteomes and gene sets of different bilaterian species in the context of the ancestral states in *Urbilateria*, it will be possible to decipher differential gene losses, gains, and duplications between these different phyla. Gene losses, gains, and duplications could then be correlated to gains, losses, and changes of biological capabilities in these lineages. These biological changes will in turn be related to environmental or geological changes at the planetary scale, as proposed by Benner *et al*. (2002). Once the ortholome (set of orthologous sequences between two or more species) has been deciphered with high reliability, reconstructions at other levels can be considered. Indeed, for example, ancestral interactomes and biological pathways could be deduced through comparisons between pathways and interaction networks from modern bilaterian species. In a similar manner, reconstruction of ancestral regulatory elements, of ancestral regulation networks, and of ancestral gene-expression patterns, could be expected as new information on expression is available for large-scale comparative studies. Thus, based upon sequence-level reconstruction, the biology of our distant ancestors can be reconstructed, and correlated with ecological and geological data. In addition to providing crucial information to understand how the genomes of bilaterian species evolved from a common ancestral genome, such information will shed light on biological mechanisms of modern species as well.

References

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., and Inoko, H. (2002) Evidence of en bloc duplication in vertebrate genomes. *Nat. Genet*. **31**: 100–105.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.

Benner, S.A., Caraco, M.D., Thomson, J.M., and Gaucher, E.A. (2002) Planetary biology–paleontological, geological, and molecular histories of life. *Science* **296**: 864–868.

Blair, J.E., Ikeo, K., Gojobori, T., and Hedges, S.B. (2002) The evolutionary position of nematodes. *BMC Evol. Biol*. **2**: 7.

Blanchette, M., Green, E.D., Miller, W., and Haussler, D. (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*. **14**: 2412–2423.

Bourque, G. and Pevzner, P.A. (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res*. **12**: 26–36.

Bourque, G., Pevzner, P.A., and Tesler, G. (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res*. **14**: 507–516.

Bourque, G., Zdobnov, E.M., Bork, P., Pevzner, P.A., and Tesler, G. (2005) Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res*. **15**: 98–110.

Castro, L.F., Furlong, R.F., and Holland, P.W. (2004) An antecedent of the MHC-linked genomic region in amphioxus. *Immunogenetics* **55**: 782–784.

Copley, R.R., Aloy, P., Russell, R.B., and Telford, M.J. (2004) Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of Caenorhabditis elegans. *Evol. Dev*. **6**: 164–169.

Danchin, E.G.J. (2004) *Reconstruction of ancestral genomic regions by comparative analysis of evolutionary conserved syntenies. Towards reconstructing the genome of the ancestor of all Bilaterian species (Urbilateria)*. PhD thesis, Aix-Marseille II, Marseilles.

Danchin, E.G.J. and Pontarotti, P. (2004a) Statistical evidence for a more than 800-million-year-old evolutionarily conserved genomic region in our genome. *J. Mol. Evol*. **59**: 587–597.

Danchin, E.G.J. and Pontarotti, P. (2004b) Towards the reconstruction of the bilaterian ancestral pre-MHC region. *Trends Genet*. **20**: 587–591.

Danchin, E.G., Abi-Rached, L., Gilles, A., and Pontarotti, P. (2003) Conservation of the MHC-like region throughout evolution. *Immunogenetics* **55**: 141–148.

Douzery, E.J., Snell, E.A., Bapteste, E., Delsuc, F., and Philippe, H. (2004) The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins hand fossils? *Proc. Natl. Acad. Sci. USA* **101**: 15386–15391.

Flajnik, M.F. and Kasahara, M. (2001) Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity* **15**: 351–362.

Gouret, P., Vitiello, V., Balandraud, N., Gilles, A., Pontarotti, P., and Danchin, E.G. J. (2005) FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinformatics* **6**: 198.

Hughes, A.L. and Friedman, R. (2004) Differential loss of ancestral gene families as a source of genomic

divergence in animals. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **271**: S107–S109.

Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., Stange-Thomann, N., Mauceli, E. *et al.* (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**: 946–957.

Jordan, I.K., Wolf, Y.I., and Koonin, E.V. (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol.* **4**: 22.

Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T., and Ishibashi, T. (1996) Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. *Proc. Natl. Acad. Sci. USA* **93**: 9096–9101.

Kasahara, M., Nakaya, J., Satta, Y., and Takahata, N. (1997) Chromosomal duplication and the emergence of the adaptive immune system. *Trends Genet.* **13**: 90–92.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**: R7.

Kunin, V. and Ouzounis, C.A. (2003) GeneTRACE-reconstruction of gene content of ancestral species. *Bioinformatics* **19**: 1412–1416.

Telford, M.J. (2004a) Animal phylogeny: back to the coelomata? *Curr. Biol.* **14**: R274–R276.

Telford, M.J. (2004b) The multimeric beta-thymosin found in nematodes and arthropods is not a synapomorphy of the Ecdysozoa. *Evol. Dev.* **6**: 90–94.

Trachtulec, Z. and Forejt, J. (2001) Synteny of orthologous genes conserved in mammals, snake, fly, nematode, and fission yeast. *Mamm. Genome* **12**: 227–231.

Trachtulec, Z., Hamvas, R.M., Forejt, J., Lehrach, H.R., Vincek, V., and Klein, J. (1997) Linkage of TATA-binding protein and proteasome subunit C5 genes in mice and humans reveals synteny conserved between mammals and invertebrates. *Genomics* **44**: 1–7.

Vienne, A., Rasmussen, J., Abi-Rached, L., Pontarotti, P., and Gilles, A. (2003a) Systematic phylogenomic evidence of en bloc duplication of the ancestral 8p11.21–8p21.3-like region. *Mol. Biol. Evol.* **20**: 1290–1298.

Vienne, A., Shiina, T., Abi-Rached, L., Danchin, E., Vitiello, V., Cartault, F., *et al.* (2003b) Evolution of the proto-MHC ancestral region: more evidence for the plesiomorphic organisation of human chromosome 9q34 region. *Immunogenetics* **55**: 429–436.

Wolf, Y.I., Rogozin, I.B., and Koonin, E.V. (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res.* **14**: 29–36.