Available online at www.sciencedirect.com
SCIENCE DIRECT®
PERGAMON
Developmental and Comparative Immunology 27 (2003) 763–779
Developmental & Comparative Immunology
www.elsevier.com/locate/devcompimm

IMGT/PhyloGene: an on-line tool for comparative analysis of immunoglobulin and T cell receptor genes

Olivier Elemento[a], Marie-Paule Lefranc[a,b,*]

[a]*IMGT, the International ImMunoGeneTics Information System®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Université Montpellier II, UPR CNRS 1142, Institut de Génétique Humaine (IGH), 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France*
[b]*Institut Universitaire de France, France*

Received 4 September 2002; revised 17 March 2003; accepted 19 March 2003

Abstract

IMGT/PhyloGene is an on-line software package for comparative analysis of immunoglobulin (IG) and T cell receptor (TR) variable genes of all vertebrate species, newly implemented in IMGT, the international ImMunoGeneTics information system ®. IMGT/PhyloGene is strongly associated with the IMGT gene and allele nomenclature and with the IMGT unique numbering for V-REGION, which directly creates standardized alignments from IMGT reference sequences. IMGT/PhyloGene is the first tool to use the IMGT expertized and standardized data for automated comparative analyses, and the first on-line software package for phylogenetic reconstruction to be integrated to a sequence database. Starting from a standardized alignment of selected sequences, IMGT/PhyloGene computes a matrix of evolutionary distances, builds a tree using the Neighbor-Joining (NJ) algorithm, and outputs various graphical tree representations. The resulting IMGT/PhyloGene tree is then used as a support for studying the evolution of particular subregions, such as the CDR-IMGT (Complementarity Determining Regions) or the V-RS (Variable gene Recombination Signals). IMGT/PhyloGene is freely available at http://imgt.cines.fr.

Keywords: Comparative analysis; Evolution; Phylogeny; Bioinformatics; Database; Immunogenetics; Immunoglobulin; T cell receptor

1. Introduction

IMGT, the international ImMunoGeneTics information system ® [1,2] (http://imgt.cines.fr), is a high quality information system specializing in immunoglobulins (IG), T cell receptors (TR) and major histocompatibily complex (MHC) molecules. In January 2003, IMGT/LIGM-DB, the IMGT comprehensive database of IG and TR annotated sequences, contained more than 67,000 sequences from 105 vertebrate species. Common access to these data through IMGT now makes it possible to perform large scale studies related to the evolution of immunoglobulin (IG) and T cell receptor (TR) genes. The IG and TR variable and constant genes

* Corresponding author. Address: IMGT, the International ImMunoGeneTics Information System®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Université Montpellier II, UPR CNRS 1142, Institut de Génétique Humaine (IGH), 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France. Tel.: +33-4-99-61-99-65; fax: +33-4-99-61-99-01.
E-mail address: lefranc@ligm.igh.cnrs.fr (M.-P. Lefranc).

0145-305X/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0145-305X(03)00078-8

are organized in clusters of duplicate genes and multigene families [3,4], assumed to have arisen through repetitive unequal crossovers. These genes have recently been the subject of a large number of evolutionary analysis, e.g. [5–7] for the IG variable genes, and [8–10] for TR variable genes. Divergent evolution and evolution by the 'birth and death' process are good candidates for explaining the evolution of these duplicate genes [5,11]. Divergent evolution (also called diversifying selection) drives the rapid differentiation of gene sequences, necessary to adapt the spectrum of the immune response to the environment. Evolution via a 'birth and death' process implies that gene duplications frequently occur in these gene clusters, and that many duplicate genes die out from deleterious mutations. Indeed, it seems that some duplicate genes are positively selected and confer advantages to some individuals, while others are not and become non-functional pseudogenes. The proportion of observed pseudogenes varies extensively between clusters within the same species, and also between species. For example, the proportion of pseudogenes within the human IG clusters is relatively large [3], but it is low within all the human TR clusters [4]. Another particular feature of these clusters is that their size (in term of number of genes) varies extensively among species [12]. Even within the same locus (such as the human TRGV cluster for example), some genes have frequently been duplicated, while some other ones have remained in single copy through evolution [13]. Comparative analyses involving gene sequences and protein structures also show that the relatively high variability at the sequence level does not prevent the IG and TR domain 3D structures from being highly conserved. All these observations indicate that many questions regarding the evolution of these genes have yet to be answered. As we show in this paper, combining bioinformatics and standardized data provides many ways and methods to study the evolution of these gene sequences.

The most insightful way to trace the evolutionary relationships between a set of gene sequences is to reconstruct their phylogenetic tree. This phylogenetic tree enables rapid and efficient visual comparison of the sequence identity levels between the different genes. It can also help to spot positively selected amino acids [14], or to detect correlated mutations at the sequence level [15,16]. However, doing phylogenetic analysis out of data coming from on-line databases often remains a daunting task, since it requires several steps: downloading the sequences, aligning them using multiple alignment softwares, visually checking the alignment, choosing a reconstruction method (among a large spectrum of available ones), selecting the appropriate software, rooting the obtained tree, evaluating the reliability of the tree, and finally drawing the final tree. Moreover, there is no standardization for the sequence selection, and different multiple alignment computer programs (CLUSTALW [17], DIALIGN [18], etc.) can be used, with different parameters (default or custom, positions with gaps removed or not), which often results in different alignments (see [19,20], for example). These different alignments can then be used with one of the numerous reconstruction methods provided in software packages such as PAUP [21], PHYLIP [22] or MEGA [23], and generally different trees are obtained. The problem is that there is no way to compare trees reconstructed from different data and with different parameters. In order to provide a common base for IG and TR sequence comparison and evolutionary analyses, we created IMGT/PhyloGene, the first tool to use the IMGT expertized and standardized sequence data for automated comparative analyses, and the first on-line software package for phylogenetic reconstruction to be integrated to a sequence database.

2. Materials

2.1. IMGT/PhyloGene standardized data

2.1.1. IMGT reference sequences

The IMGT/PhyloGene standardized reference sequence data consists of the V-REGION alleles (*01) from the IMGT reference directory sets which comprise one representative of each functional or ORF allele of each gene (available in IMGT Repertoire [3,4,24], http://imgt.cines.fr). In IMGT/PhyloGene, as in the other IMGT databases, Web resources and tools, genes are named according to the IMGT nomenclature. IMGT genes and alleles names were approved by the HUGO Nomenclature

Committee (HGNC) in 1999 [25] and entered in LocusLink (NCBI, USA), GDB (Toronto, Canada) and GeneCards [3,4,26,27].

2.1.2. IMGT unique numbering

In the IMGT reference sequences, gaps and delimitations of the framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT) are set according to the IMGT unique numbering for V-REGION [3,4,28–30]. The IMGT unique numbering for V-REGION is valid whatever the receptor type (IG or TR), whatever the chain type (heavy, kappa, lambda for the IG; alpha, beta, gamma, delta for the TR) and whatever the species. It is also valid for the V-LIKE domains of non-IG or non-TR molecules [28–30].

2.1.3. Benefits of using standardized data

IMGT/PhyloGene standardized sequences obviate the need for very time consuming sequence selection of previously known sequences and multiple alignment procedures. They provide a reliable and controlled set of reference sequences, allowing trees to be compared. Moreover, all the sequences respect the same common rules, since they are aligned according to the IMGT unique numbering. This means that conserved amino acids are at the same standardized position, and that gaps are inserted at the correct standardized positions. In this situation, a set of sequences constitutes a multiple alignment in itself.

2.1.4. IMGT/PhyloGene database

For a faster access, the V-REGION allele (*01) sequences used for analysis are stored within a separate database. The IMGT/PhyloGene database also contains the corresponding V-RS sequences, since the software offers the possibility to use the reconstructed tree as support for evolutionary analyses of the V-RS sequences. The IMGT/Phylogene database currently contains 724 IG and TR alleles (*01), which represent one sequence for each functional or ORF IG and TR V-GENE from *Homo sapiens* and *Mus musculus* (Table 1) [3,4].

2.1.5. User-entered data

IMGT/PhyloGene can analyse V-REGION sequences which are not contained in its database but which are provided by the user. The only

Table 1
Content of the IMGT/PhyloGene database for the *H. sapiens* and *M. musculus* V-REGION alleles (*01) and availability of the other alleles

Species	Group	Number of alleles (*01)[a]	Number of all alleles[b]	Accession number[c]	References
H. sapiens	IGHV	55	237	ALIGN_000299	[3,31–33]
H. sapiens	IGKV	45	64	ALIGN_000321	[3,33–35]
H. sapiens	IGLV	39	78	ALIGN_000420	[3,33,36,37]
H. sapiens	TRAV	46	104	ALIGN_000306	[4,38,39]
H. sapiens	TRBV	64	138	ALIGN_000302	[4,39,40]
H. sapiens	TRDV	3	6	ALIGN_000305	[4,39]
H. sapiens	TRGV	10	19	ALIGN_000191	[4,39]
M. musculus	IGHV	211	234	ALIGN_000439	
M. musculus	IGKV	119	131	ALIGN_000434	[41]
M. musculus	IGLV	3	5	ALIGN_000435	
M. musculus	TRAV	86	209	ALIGN_000445	[83]
M. musculus	TRBV	26	52	ALIGN_000440	[42]
M. musculus	TRDV	10	16	ALIGN_000443	[43,83]
M. musculus	TRGV	7	27	ALIGN_000444	
Total		724			

Sequences in FASTA format and with gaps according to the IMGT numbering are available in IMGT Repertoire, http://imgt.cines.fr. Alignments in CLUSTALW format and IMGT numbering are available in IMGT-Align, http://imgt.cines.fr (IMGT Repertoire > Proteins and alleles), and in EMBL-Align [44] via SRS (http://srs.ebi.ac.uk). Locus representations, germline gene tables, alignments of alleles and Colliers de Perles [3,4,24,45] are available in IMGT Repertoire.

[a] Number of V-REGION allele (*01) sequences, per group, in the IMGT/PhyloGene database.

[b] Number of all V-REGION allele sequences, per group, in the IMGT-and EMBL-Align alignments.

[c] The accession numbers are identical for IMGT-Align and EMBL-Align.

constraint is that these sequences are entered in FASTA format and with gaps according to the IMGT numbering. This format with gaps can be obtained using IMGT/V-QUEST [1,45–47] (http://imgt.cines.fr). If other V-REGION alleles, different from (*01), need to be entered, then they can be retrieved from the IMGT reference directory V-REGION sets or from IMGT-or EMBL-ALIGN (Table 1).

3. Methods

3.1. Choice of the distance approach for reconstructing phylogenies in IMGT/PhyloGene

There are many different methods to build a phylogenetic tree from sequence data (see [48] for a detailed review of the different methods):

(1) *Parsimony methods*. Given a set of observed homologous sequences, the goal of parsimony methods is to find the shortest tree in terms of number of mutations required to obtain these sequences. The inherent simplicity of parsimony methods has always made them attractive. However, whereas they usually behave quite well with sequences presenting a high rate of identity among themselves, it has been shown that parsimony methods suffer from several drawbacks [49]. One of these drawbacks is their statistical inconsistency, which implies that the probability for them to find the correct tree (i.e. the real history) does not always increase as the length of the sequence increases. A consequence of this inconsistency is that parsimony methods are sensible to the long branch attraction phenomenon [49]. Indeed, if two branches in the real tree are very long, they might incorrectly end up joined together in the reconstructed tree. This is caused by superimposed changes, which make highly divergent sequences look closer than they actually are.

(2) *Maximum likelihood methods* [50]. They are assumed to be the most accurate reconstruction methods. In particular, they have been shown to be relatively robust of violations of the models of evolution they usually rely on [51]. Since they use computationally heavy optimization procedures, the main drawback with maximum likelihood methods is that they are very slow.

(3) *Distance methods*. They take as input a set of pairwise distances between sequences, estimated from nucleotide or protein sequences according to a selected model of evolution. The use of a model of evolution enables to correct the observed distances for surimposed changes, and therefore limits the effect of the long branch attraction phenomenon. Their main drawback is that reducing a multiple alignment to a distance matrix necessarily results in a loss of information, which may limit the topological accuracy of the results. The main benefit of distance methods is that they are generally fast, and they enable the use of sophisticated models of evolution (such as those taking into account heterogeneous rates of substitution among sites, for example).

In IMGT/PhyloGene, we adopted a distance approach, with the main goal of maximizing the speed of the on-line analysis. Phylogenetic analysis using a distance approach necessitates three distinct steps: (a) aligning the sequences, such that each position in the multiple alignment corresponds to the same position in the ancestral sequence. As shown above, the alignment step is not necessary when using IMGT/PhyloGene; (b) computing a distance matrix between the sequences; and (c) reconstructing a tree from this distance matrix.

3.2. Computing a distance matrix: choice of models and parameters in IMGT/PhyloGene

In the context of phylogenetic analyses, a distance matrix is simply a square matrix in which each entry corresponds to an estimated evolutionary distance between two sequences from a multiple alignment. In its simplest form, an evolutionary distance between two sequences can be defined as the proportion of differences, at the nucleotide level, between the two sequences. The proportion of differences is the ratio between the number of nucleotide substitutions necessary to transform one sequence into the other and the length of the aligned sequences.

Since there can be hidden substitutions in the course of evolution, the proportion of observed differences between two sequences often underestimates the real number of substitutions. For this reason, the observed distance between the two sequences is corrected using one of the available stochastic models of evolution (see [48] for a review). We considered four distinct models of evolution for use with IMGT/PhyloGene: Jukes-Cantor (JC69) [52], Kimura 2-parameters (K2P) [53],

F84 [54,55], and the Poisson correction [48]. The JC model assumes equal rates of substitutions among all nucleotides, while the K2P model differentiates the rate of transitions from the rate of transversions. The F84 model is similar to the K2P model, except that it assumes unequal nucleotide frequencies. The Poisson correction model is an amino acid distance, and assumes that the number of amino acid substitutions at each site follows the Poisson distribution. Using an amino acid distance can be useful when analysing coding sequences, since synonymous sites mutate faster than the other sites, and are very often saturated.

All of these models assume that the substitution rates are equal along the multiple alignment, which may not always be the case [56]. The rate heterogeneity within a set of aligned sequences is generally modeled using a gamma distribution. A unique parameter, denoted α, determines the shape of the distribution and therefore the level of rate heterogeneity within the sequences. When $\alpha < 1$, the gamma distribution is close to an exponential distribution, which implies that the rate heterogeneity is very important. When α grows, the gamma distribution becomes closer and closer to a normal distribution, and tends towards rate homogeneity. For a given set of sequences, the value of the α parameter can be estimated using maximum likelihood methods (e.g. PAML [57]).

To assess the different models of evolution in the IMGT/PhyloGene context, we constructed several datasets of IG and TR sequences, as representatives of different analyses which could be performed with IMGT/PhyloGene. The first datasets contained sequences from the same group and same species, but from different subgroups. The second datasets contained sequences from the same species, but from different groups. The third datasets contained various sequences from different species. For each dataset, we calculated the distance matrices corresponding to the above models. Then, we applied Neighbor-Joining (NJ) (see below) with each of the distance matrices as input and compared the obtained trees. We also used PAML [57] to estimate both the α parameter and the transition/transversion ratio, for each dataset. We then used the obtained values to create refined trees, and compared those trees to the initial ones.

For each dataset, the initial trees, obtained with the JC, K2P and F84 models, were almost identical, in terms of topology. The trees obtained with the Poisson distance were slightly different from the trees obtained with the other models. However, the group and subgroup distinctions were conserved. The average estimated α parameter was 1.296, while the average estimated transition/transversion ratio was 2.343, over all datasets. Using these parameters, new distances matrices were re-estimated using DNADIST, from the PHYLIP package, and the corresponding trees were reconstructed using NJ. For all datasets, the topology of the refined trees were identical to the topology of the initial trees, except for a few very short branches.

These results tend to show that taking into account rate heterogeneity and the estimated transition/transversion ratio does not provide visible differences or improvements to the reconstructed trees, in the context of the IMGT/PhyloGene analyses. While it is likely that they could improve the results in some cases, the α parameter and the transition/transversion ratio seem very often dataset-specific. Unfortunately, estimating these parameters would not be possible in the IMGT/PhyloGene context, due to the enormous computational burden of the estimation process.

For these reasons, we decide to use only two models of evolution in IMGT/PhyloGene: the F84 model and the Poisson correction. The F84 model is the most complex and generic model, among those available. The Poisson correction is well suited when the synonymous substitutions rate is well higher than the non-synonymous substitutions rate. For the above reasons, we do not use a gamma correction in those models. We also do not make any particular assumptions regarding the expected transition/transversion ratio and use the distance equations provided in Ref. [54,48].

When computing the distance matrix, IMGT/PhyloGene discards the highly variable positions corresponding to the CDR1-IMGT, CDR2-IMGT and CDR3-IMGT, by default. This strategy is generally adopted for the IG and TR variable genes [5,6,8]. Indeed, CDR regions evolve not only by substitutions, but also by codon insertions and deletions, and are difficult to align with accuracy, due to their unequal lengths and high variability. Since the IMGT numbering provides information about the position and length of these regions, it is very easy to discard them, in a standardized way, in the IMGT/PhyloGene

tool. However, IMGT/PhyloGene has an option which allows to keep the CDR-IMGT in the analysis, and which can be used, for example, when the V-REGIONs have identical CDR-IMGT lengths. This option is particularly useful for comparative analysis example of mutated variable sequences during the antibody affinity maturation (in vivo or in vitro). In both cases (default or option), the CDR-IMGT are stored in the IMGT/PhyloGene database and, as shown below, can be displayed later at the tips of the obtained tree.

3.3. Synonymous and non-synonymous substitution rates in IMGT/PhyloGene

In IMGT/PhyloGene, the synonymous and non-synonymous substitution rates between two nucleotide sequences are estimated using the Gojobori and Nei method [58]. Comparing synonymous and non-synonymous substitution rates allows to gain better insight into the evolutionary pressure acting at the sequence level [5,6,11]. Given a group of homologous sequences, it is generally considered that when the average synonymous substitution rate (K_s) is greater than the average non-synonymous substitution rate (K_a), the sequences are undergoing 'purifying' selection, which means that non-synonymous mutations are eliminated. On the opposite side, when K_s is smaller than K_a, the sequence are undergoing 'diversifying' selection, which means that non-synonymous mutations are favoured.

3.4. Tree reconstruction: choice of the Neighbor-Joining method in IMGT/PhyloGene

In IMGT/PhyloGene, phylogenetic trees are reconstructed using the NJ algorithm [59]. The NJ algorithm is one of the most popular methods for reconstructing phylogenetic trees from a matrix of pairwise evolutionary distances. Starting from a star tree (in which all the taxa are connected to a single interior node), the NJ algorithm follows an agglomerative scheme: it iteratively picks a pair of taxa, creates a new node which represents the cluster of these taxa, estimates the branch length between the two taxa and the new node, and reduces the distance matrix by replacing both taxa by this node. This cycle is repeated until only three taxa remain. To agglomerate pair of nodes, NJ follows the minimum-evolution principle, which consists in selecting the tree with the smallest sum of branch lengths. It must be noticed that NJ does not always find the shortest possible tree, since the agglomeration process is guided by a greedy heuristics. However, this does not prevent NJ from showing good performances, because the correct tree itself is generally not the shortest one, but only close to the shortest one [60]. Besides its good performances, NJ is one of the fastest tree reconstruction methods available today, and is therefore very well suited to on-line analysis.

3.5. Rooting the trees: midpoint and outgroup methods used in IMGT/PhyloGene

Rooted trees are much more understandable than unrooted trees, since the labels (i.e. the gene names) are regularly spaced on the vertical axis. However, the tree reconstruction algorithm (NJ) used in IMGT/PhyloGene yields unrooted trees, like most phylogenetic reconstruction methods. The reason is that the most frequently used models of evolution (such as the K2P model [53]), which form the basis of many reconstruction algorithms (such as NJ), are reversible and do not define an evolutionary direction (for example, an 'A' is free to mutate to a 'G' and then back to an 'A'). As a consequence, the location of the root within the tree cannot be determined. Therefore, rooting a phylogenetic tree constitutes an additional procedure, which is generally performed after the reconstruction procedure. There are two main methods for rooting the trees [48], which are both implemented in IMGT/PhyloGene: the midpoint method and the outgroup method.

The midpoint method assumes a weak form of molecular clock mode of evolution: assuming that the most divergent lineages have evolved at the same rate, the midpoint method locates the root at the midpoint of the tree path connecting the two most divergent taxa. In IMGT/PhyloGene, the trees built using the NJ algorithm are rooted by default using the midpoint method.

The outgroup method involves selecting one or several outgroup sequences. These sequences, which are assumed to lie outside the monophyletic group of interest, need to be included in the analysis. After reconstruction, the position in the tree where the outgroup branches to the monophyletic group of interest locates the root of the tree. The choice of

the outgroup is obviously very important and depends on the sequences within the analysis. A set of IG V-REGIONs can be rooted using one or several TR V-REGIONS, and conversely, a set of TR V-REGIONs can be rooted using one or several IG V-REGIONs. A set which would include IG and TR V-REGIONs can be rooted using outgroup sequences from the V-LIKE immunoglobulin superfamily [30,61] sequences, such as CD4, CD8A and CD8B. The use of several outgroup sequences instead of a single one is assumed to provide a more accurate rooting, because it reduces the long branch attraction phenomenon. Moreover, rooting the tree with one or several outgroup sequences is often assumed to be more reliable than using the midpoint procedure.

3.6. Drawing trees with or without branch lengths: IMGT/PhyloGene implementation

In IMGT/PhyloGene, the rooted trees can be drawn with or without branch lengths. In a tree drawn 'with branch lengths', branch lengths have a meaning in term of evolutionary distances between genes (ancestral and observed). In a tree drawn 'without branch lengths', branch lengths do not have any particular meaning in term of evolutionary distances, and are just used to provide a horizontal alignment of the gene names. Although there are several software packages for drawing trees, e.g. DRAWGRAM, DRAWTREE, from the PHYLIP package [22], NJPLOT [62], or PHYLODENDRON [63], none of them was fast and flexible enough to be integrated to an on-line tree reconstruction and visualisation tool. Consequently, we implemented our own tree visualisation application for drawing rooted phylogenetic trees, which produces vertically oriented phenograms out of rooted trees, with or without branch lengths and in a format suitable for Web display.

3.7. Original features of the IMGT/PhyloGene trees

Once a IMGT/PhyloGene tree has been obtained, it is possible to use it as a support for further evolutionary analyses. For example, it is possible to study the evolution of certain subsequences by displaying them at the tips of the rooted phylogenetic tree, along with the corresponding gene names, accession numbers and species. The tree must be drawn without branch lengths, so as to align the subsequences. In IMGT/PhyloGene, it is possible to display the complementarity determining regions (CDR1-IMGT, CDR2-IMGT and germline CDR3-IMGT), the framework regions (FR1-IMGT, FR2-IMGT and FR3-IMGT) or, when available, the variable recombination signals (V-RS).

This feature is particularly useful for CDR-IMGT, since it will give clues about the evolution of this crucial antigen binding region. For example, the number of amino acids that constitute the CDR is important for the variability and capacity to bind antigens [64] or to get insights into the evolution of the IG [7,65]. The CDR-IMGT lengths are crucial information to characterise subgroups and genes [3,4,66]. For that reason, a procedure has been added in IMGT/PhyloGene, which allows to reconstruct the ancestral CDR lengths (in number of nucleotides) using a Most Parsimonious Reconstruction (MPR) algorithm [48,67].

The possibility given by IMGT/PhyloGene to surimpose the V-RS sequences with the tree calculated on the framework regions may also prove very useful for understanding the evolution of these sequences [68] and the mechanisms involved in V-D-J rearrangements.

3.8. IMGT/PhyloGene implementation

IMGT/PhyloGene is fully integrated to the rest of the IMGT information system. The sequences are stored in a MySQL relational database [69], so as to increase the speed of the data retrieving operations. The MySQL database is regularly updated to include newly added reference sequences, or to mirror the changes made to some reference sequences within the IMGT reference directory sets. The distance calculation procedure, the tree building algorithm, the rooting procedures, the generation of graphical tree representations and the most parsimonious reconstruction of CDR-IMGT lengths have been implemented in the C language [70]. The estimation of synonymous and non-synonymous substitution rates, and the scripts which generate Web pages have been implemented in the Perl language [71].

4. Results

4.1. IMGT/PhyloGene selection page

The first Web page in IMGT/PhyloGene is the sequence selection page (Fig. 1). This page allows to select V-REGION gene sequences from the IMGT/PhyloGene database, and also to add user-supplied V-REGION gene sequences to the analysis. Due to limitations of HTML-based user interfaces, the selection process is still the most time-consuming task in IMGT/PhyloGene. However, once the selection is done, it takes less than a minute to reconstruct a phylogenetic tree with IMGT/PhyloGene and the user is only two clicks away from obtaining a graphical tree representation out of the selected sequences.

The IMGT/PhyloGene selection process follows an interactive approach and can be seen as filling a bag (or a cart) of sequences. When starting an IMGT-PhyloGene session, the user's selection is empty. As shown in Fig. 1, three drop-down menus on the left part of the screen allow the user to select the desired species, group, and (optionally) subgroup, as described in IMGT-ONTOLOGY [3,4,72–74]. This will automatically reload the screen with the corresponding gene names displayed in the multiple selection list, on the right part of the screen. The user can then select the genes to analyse from the list. Then, clicking on the 'Add to selection' button will add the corresponding sequences to the selection. At any moment, the user can empty the selection using the 'Reset' button.

On the same Web page, a text area allows the user to supply its own sequences, in FASTA format. To be compared with the other sequences, the user-supplied

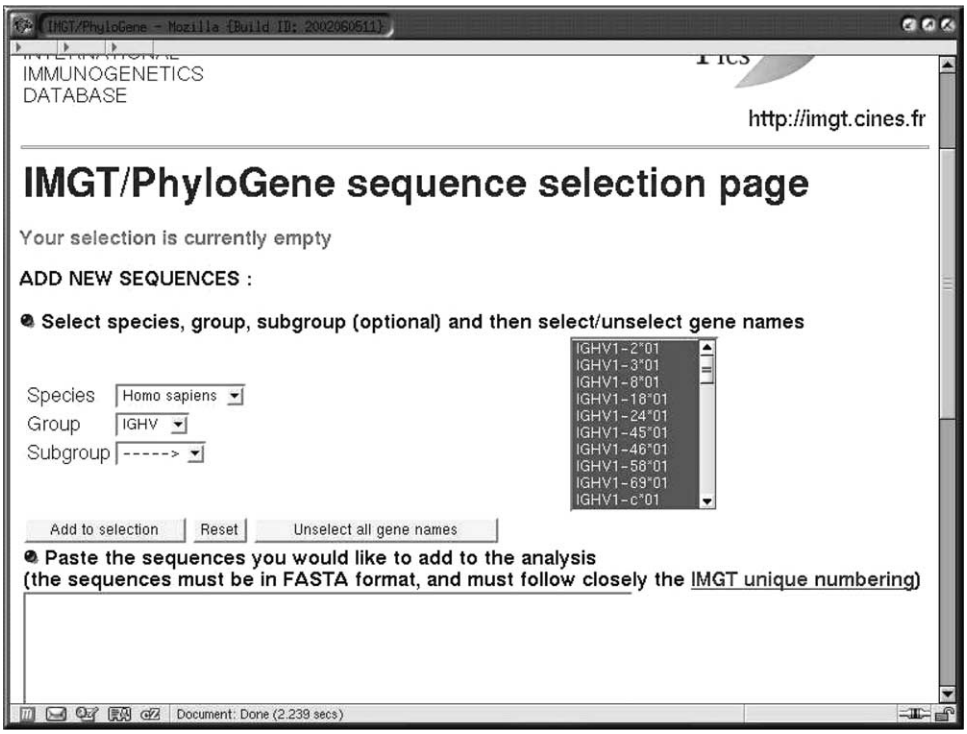


Fig. 1. IMGT/PhyloGene sequence selection page. The three drop-down menus on the left part of the screen allow the user to select the desired species, group and subgroup. This will automatically reload the screen with the corresponding gene names, displayed in the multiple selection list on the right part of the screen. For example, for a selection of the *H. sapiens* IGHV group and IGHV1 subgroup, the list of the IGHV1 gene names will automatically appear in the gene name selection list. The user can then select the gene sequences to analyse (by default, all the genes are selected). The corresponding sequences are those of the allele (*01) of each selected gene. The user can add his (her) own sequences by pasting them in the text area located in the lower part of the screen (for more information, refer to the text and to the IMGT/PhyloGene Documentation on the IMGT web site, http://imgt.cines.fr).

sequences need to respect both the maximum length (in number of nucleotides) of the V-REGION and the gap positions of the FR-IMGT and CDR-IMGT according to the IMGT unique numbering. The IMGT/V-QUEST tool [46,47], freely available online at http://imgt.cines.fr, provides V-REGION sequences in FASTA format with gaps according to the IMGT unique numbering, ready for use in IMGT/PhyloGene. However partial sequences, or sequences which are too distant from those of the IMGT reference directory may be incorrectly aligned by IMGT/V-QUEST. In this case, the positioning of the gaps should be done manually, and partial sequences in 5′ and/or 3′ should be completed with gaps.

If the tree has to be rooted with one or several outgroup sequences, the gene name(s) of the sequence(s) can be selected from the IMGT/PhyloGene multiple selection list, or the user outgroup sequence(s) can be pasted into the text area, at this step.

4.2. IMGT/PhyloGene selected sequences page

Each time a set of sequences is added to the user's selection, the whole resulting selection is displayed in the IMGT/PhyloGene selected sequences page (Fig. 2). From this page, the user can perform the following actions: (1) go back to the selection page, so as to select more sequences; (2) compute a distance matrix from the selected sequences; (3) compute the synonymous and non-synonymous substitution rates within the selected sequences. From the same Web page, the user can also download the multiple alignment in the standard non-interleaved PHYLIP format to his (her) own computer. This feature can be useful if using another phylogenetic reconstruction software is needed. The F84 model was chosen as the default option for distance matrix computation, because it is a reasonably complex model for nucleotide sequences. However, if the estimated synonymous substitutions rate is

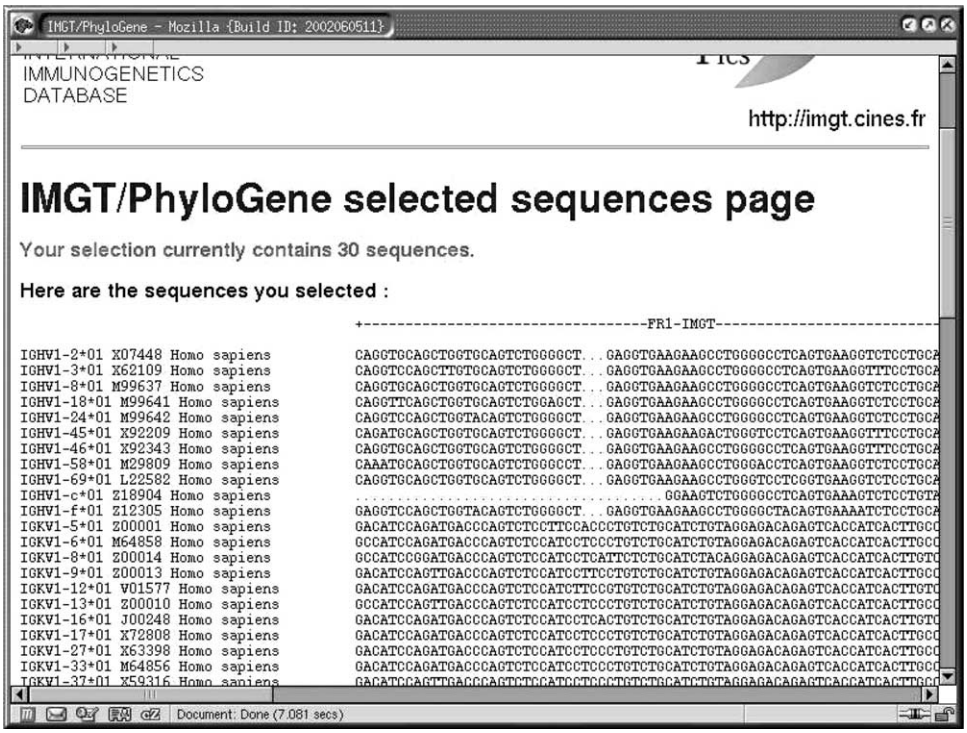


Fig. 2. IMGT/PhyloGene selected sequences page. On this screen, the selected sequences are aligned, and the regions corresponding to the FR-IMGT and CDR-IMGT are shown. Note that the CDR-IMGT are automatically removed from the alignments for the phylogenetic analysis, but can still be displayed as an output option of the resulting trees.

significantly higher than the non-synonymous substitutions rate, the Poisson model is likely to provide more accurate estimates of the evolutionary distances between the selected sequences. However, the Poisson distance requires the nucleotide sequences to be translated into amino acid sequences, which decreases the sequence length and therefore increases the variance of the estimates.

4.3. IMGT/PhyloGene substitution rates result page

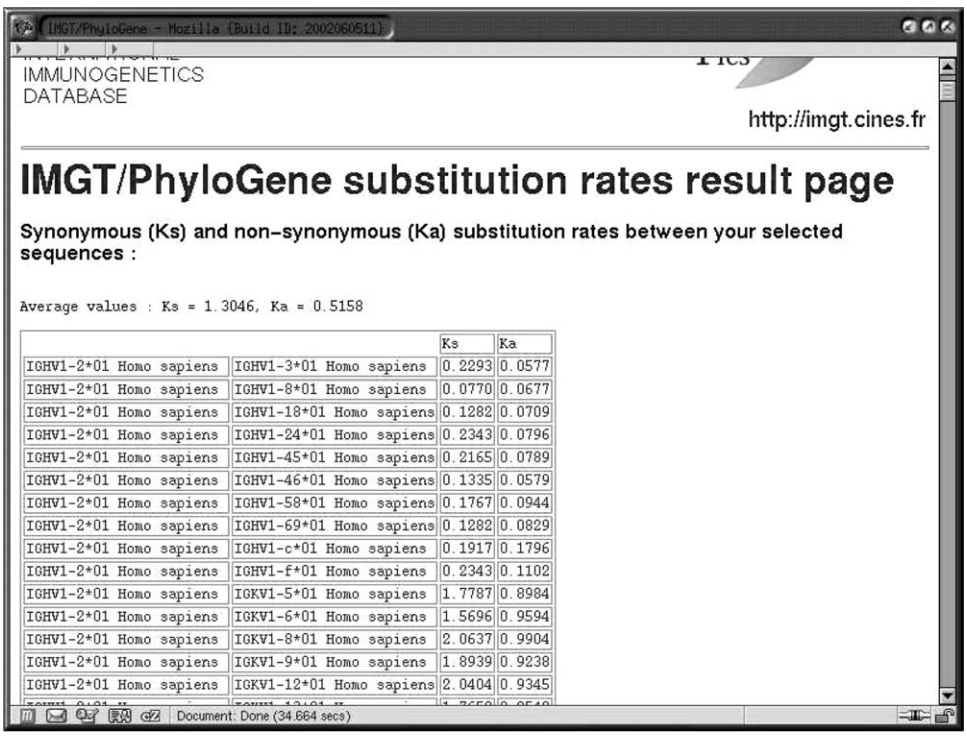
The IMGT/PhyloGene substitution rates result page (Fig. 3) provides, at the top of the page, the average values of K_s and K_a for all the analysed sequences. The K_s and K_a for the sequences compared two by two are displayed in the table below these average values. When the K_s is superior to the K_a (denoting 'purifying selection'), the row of the array is colored in blue. When the K_s is lesser than the K_a (denoting "diversifying selection"), the row is colored in red. Note that the displayed substitution rates are corrected using a Jukes-Cantor evolution model [52]. According to this model, the correction is not applicable if the observed substitution rate is greater than 0.75. When one of the two substitution rates cannot be calculated, the row is left in black.

4.4. IMGT/PhyloGene distance matrix result page

The IMGT/PhyloGene distance matrix result page (Fig. 4) is only displayed for information purpose. Clicking on the unique Web button at the bottom of this page triggers the tree reconstruction using the Neighbor-Joining (NJ) algorithm.

4.5. IMGT/PhyloGene tree with branch lengths

The default output of IMGT/PhyloGene is a tree with branch lengths, rooted using the midpoint procedure (Fig. 5A). After selection of the outgroup sequence(s) among the analysed sequences, clicking on the 'Go' button displays the corresponding



		Ks	Ka
IGHV1-2*01 Homo sapiens	IGHV1-3*01 Homo sapiens	0.2293	0.0577
IGHV1-2*01 Homo sapiens	IGHV1-8*01 Homo sapiens	0.0770	0.0677
IGHV1-2*01 Homo sapiens	IGHV1-18*01 Homo sapiens	0.1282	0.0709
IGHV1-2*01 Homo sapiens	IGHV1-24*01 Homo sapiens	0.2343	0.0796
IGHV1-2*01 Homo sapiens	IGHV1-45*01 Homo sapiens	0.2165	0.0789
IGHV1-2*01 Homo sapiens	IGHV1-46*01 Homo sapiens	0.1335	0.0579
IGHV1-2*01 Homo sapiens	IGHV1-58*01 Homo sapiens	0.1767	0.0944
IGHV1-2*01 Homo sapiens	IGHV1-69*01 Homo sapiens	0.1282	0.0829
IGHV1-2*01 Homo sapiens	IGHV1-c*01 Homo sapiens	0.1917	0.1796
IGHV1-2*01 Homo sapiens	IGHV1-f*01 Homo sapiens	0.2343	0.1102
IGHV1-2*01 Homo sapiens	IGKV1-5*01 Homo sapiens	1.7787	0.8984
IGHV1-2*01 Homo sapiens	IGKV1-6*01 Homo sapiens	1.5696	0.9594
IGHV1-2*01 Homo sapiens	IGKV1-8*01 Homo sapiens	2.0637	0.9904
IGHV1-2*01 Homo sapiens	IGKV1-9*01 Homo sapiens	1.8939	0.9238
IGHV1-2*01 Homo sapiens	IGKV1-12*01 Homo sapiens	2.0404	0.9345

Fig. 3. IMGT/PhyloGene substitution rates result page. The average values for the K_s and K_a of all the analysed sequences are shown at the top of the page. The K_s and K_a between the sequences, compared two by two, are displayed in the table below.

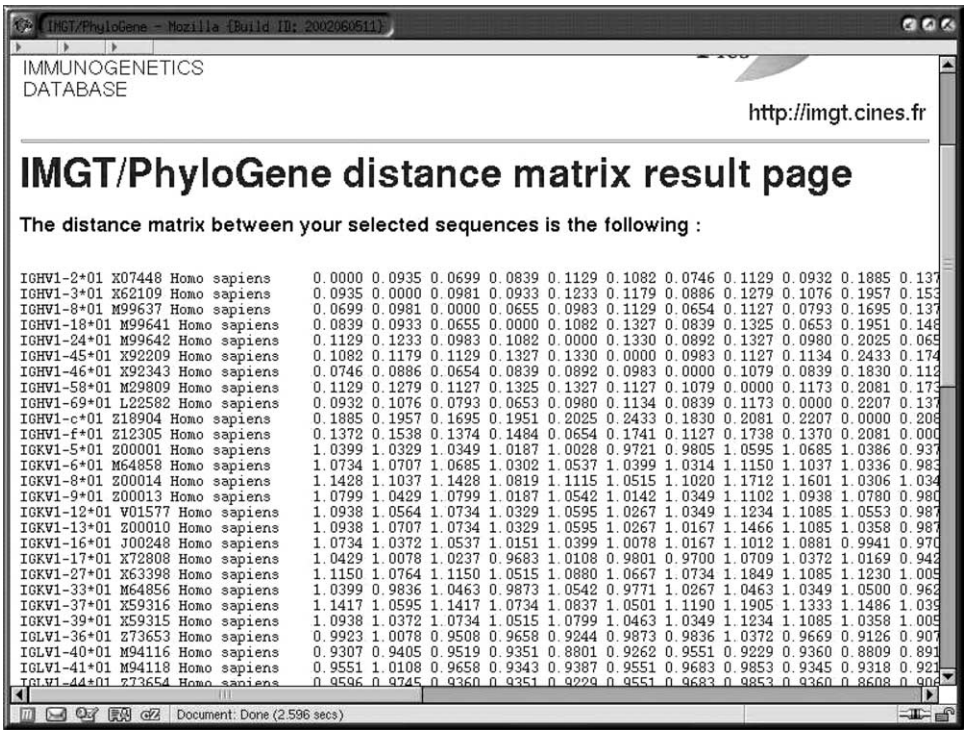


Fig. 4. IMGT/PhyloGene distance matrix result page. Clicking on the unique button at the bottom of this page (not shown) triggers the tree reconstruction using the Neighbor-Joining (NJ) algorithm.

rooted tree (Fig. 5B). At any moment, the user can re-display the same tree, rooted with the midpoint procedure. The same Web page offers two additional actions: (1) the user can go back to the sequence selection page to add outgroup sequences if needed, and (2) can download the tree in the standard Newick format. This provides a way, for example, to draw the reconstructed tree using third-party softwares.

4.6. IMGT/PhyloGene tree without branch lengths and IMGT/PhyloGene tree alignment pages

IMGT/PhyloGene trees without branch lengths (Fig. 6A) can be displayed by clicking on the IMGT/PhyloGene 'tree without branch lengths' button at the bottom of each page displaying a tree with branch lengths (and conversely). The IMGT/PhyloGene tree without branch lengths allows sophisticated displays, designated as IMGT/PhyloGene tree alignment pages (Fig. 6B). Indeed, they comprise the display of the reconstructed tree, either with the V-REGION subsequences, such as the CDR1-, CDR2-, CDR3-, FR1-, FR2-or FR3-IMGT, or the associated V-RS subsequences. To generate such a representation, the user simply needs to select the subsequences to display and then to click on the 'Show' button. If a tree with CDR1-, CDR2-or CDR3-IMGT is displayed, it is also possible to calculate and display the ancestral lengths of the observed CDR1-, CDR2-or CDR3-IMGT (Fig. 6B) by clicking on the single button at the bottom of the page.

5. Discussion and directions for further research and developments

Owing to the IMGT/PhyloGene rapidity and scalability, it is possible to build trees out of several hundreds of sequences. By analysing user sequences, together with sequences from the IMGT/PhyloGene database, the IMGT/PhyloGene tool is particularly

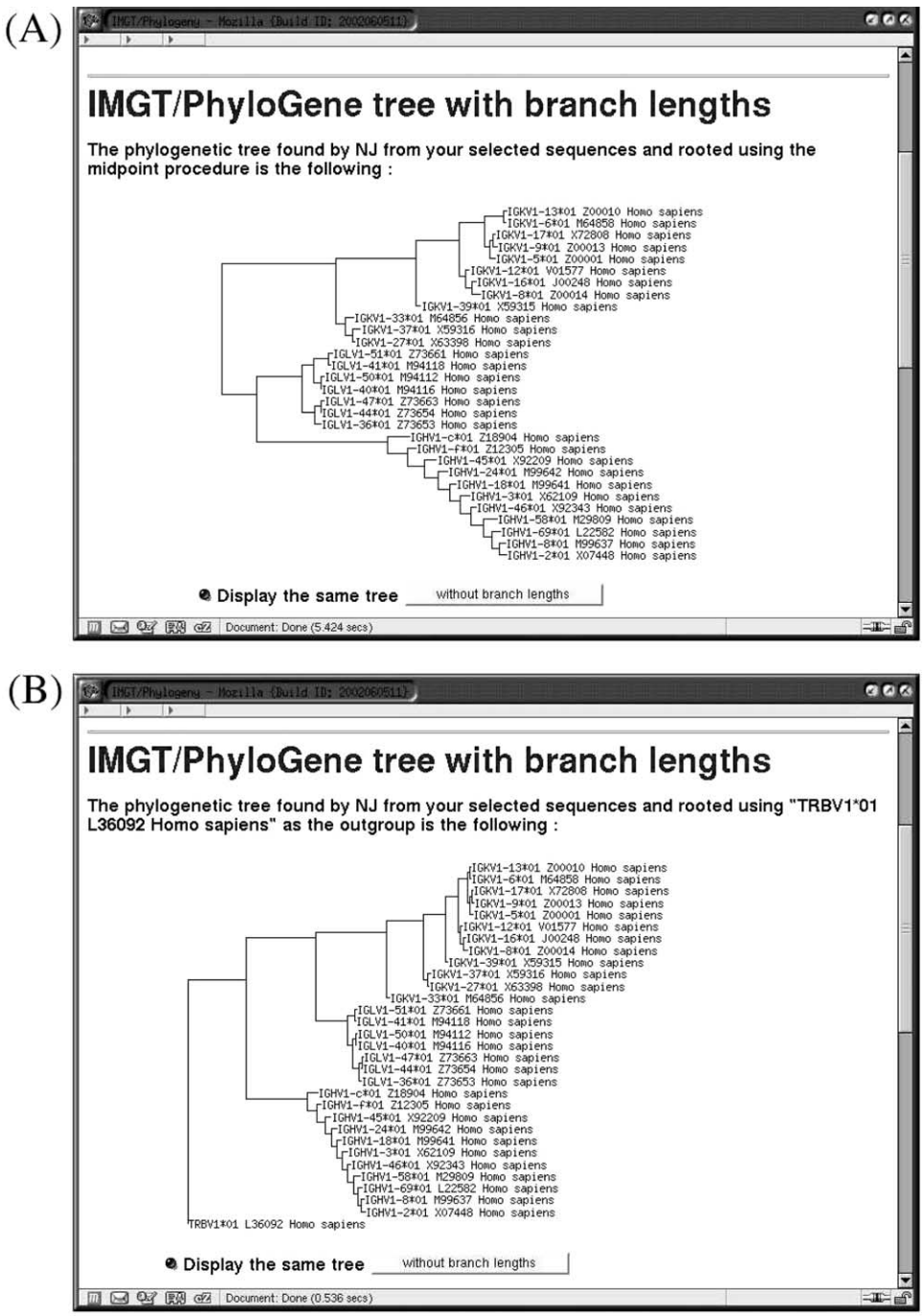


Fig. 5. IMGT/PhyloGene phylogenetic tree with branch lengths. The tree found by the Neighbor-Joining (NJ) algorithm is displayed rooted using (A) the midpoint procedure, (B) a single outgroup sequence (TRBV1*01). IMGT gene and allele name, IMGT/LIGM-DB accession number, and species latine name are displayed systematically for sequences coming from the IMGT/PhyloGene database.

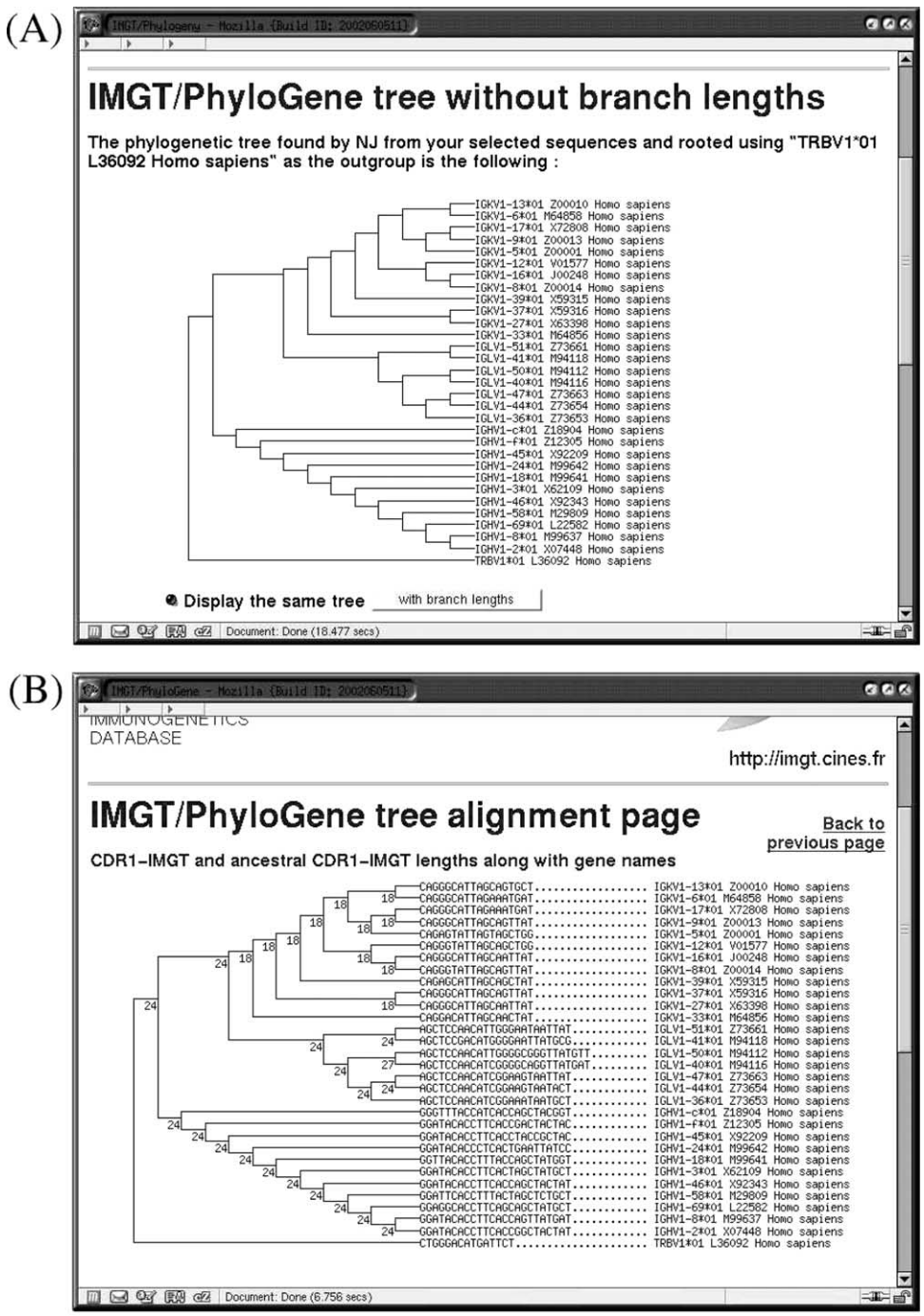


Fig. 6. IMGT/PhyloGene phylogenetic tree without branch lengths. (A) The tree found by Neighbor-Joining (NJ) algorithm is displayed rooted using a single outgroup sequence (TRBV1*01). (B) IMGT/PhyloGene tree alignment page displaying the same tree with the corresponding CDR1-IMGT subsequences at the tips of the tree. The ancestral lengths of the CDR1-IMGT, calculated using a most-parsimonious reconstruction procedure, are also displayed. IMGT gene and allele name, IMGT/LIGM-DB accession number, and species Latin names are displayed systematically for sequences coming from the IMGT/PhyloGene database.

useful to conduct sequence identity searches, to assign new genes or alleles to a given subgroup, or to relate genes from different species to human or mouse genes or subgroups. It can also be used to identify suitable human FR regions for creating humanized antibody, when using the CDR-grafting technique. Indeed, it has been shown that grafting mouse CDR onto germline human FR regions which are very close to the initial mouse FR region yields functionally equivalent humanized antibodies, with reduced immunogenicity [75–78].

IMGT/PhyloGene also provides a very useful tool for studying the evolution of the CDR-IMGT lengths and sequences, and for validating the reconstructed phylogeny. Indeed, the history of codon insertions/deletions in the CDR-IMGT and the comparison of the CDR-IMGT sequence data will allow critical assessment and validation of the phylogeny reconstructed uniquely from the framework regions. As an example, it will be of interest to check whether two insertions observed in close but concurrent branches of the reconstructed tree arose from independent insertion events (in which case the phylogeny is correct), or from a unique event (in which case the phylogeny is locally incorrect). The ability of IMGT/PhyloGene to provide insights into the evolution of V-RS sequences is also very useful, and we are currently considering extensions of this functionality to 5′ introns, leaders and other subsequences.

The only limitation with IMGT/PhyloGene is that it does not provide information related to the reliability of the reconstructed trees. The estimation is generally done using the bootstrap procedure [79], which provides numerical measures of the robustness of a reconstructed history with regard to sampling noise. Performing a bootstrap analysis consists in generating hundreds of replicated, slightly modified, multiple alignments from the initial alignment, and then in applying the same reconstruction algorithm to each of these replicate datasets. This procedure is computationally heavy and therefore not suited to on-line analysis. However there are other ways to measure the reliability of reconstructed trees, such as the interior branch tests [80]. Integrating interior branch tests within the IMGT/PhyloGene tree representations represents a future direction for development. However, since both the data used in the reconstruction and the reconstructed tree can be downloaded to the user's computer, performing bootstrap analysis using third-party softwares, such as PHYLIP, should be relatively straightforward.

At the moment, only the (*01) alleles are available in IMGT/PhyloGene. Introducing the other alleles within the database and allowing the user to select them without flooding him with data represents a direction for further research. Nonetheless, allele sequences can already be included into the analysis using the text area at the bottom of the selection page. However, it is important to remember that the default option in IMGT/PhyloGene automatically discards the CDR-IMGT regions for the phylogenetic analysis. In this case, the variability located in the CDR-IMGT is not taken into account, and allele sequences may end up incorrectly placed within the resulting tree. Also note that the nucleotide and amino acid differences between two alleles can be visualised in the IMGT Alignments of alleles (in the IMGT repertoire), and analysed with the IMGT/Allele-Align tool, also available at http://imgt.cines.fr. Another development direction is to add pseudogenes to the sequences available for selection. The main problem is that many of them are partial and/or too divergent to position gaps according to the IMGT unique numbering.

It must be noted that the NJ method is usually used to reconstruct speciation trees, i.e. trees aimed at representing the relationships between sets of observed species. In the IMGT/PhyloGene context, the goal is to rapidly reconstruct gene trees for sets of IG and TR genes. While gene trees are conceptually different from speciation trees, they can be reconstructed using the same methods and algorithms as those used in reconstructing speciation trees from sequences. Nonetheless, there exists ongoing research to integrate the information provided by the gene order on the chromosome to get deeper insights into the evolution of gene families [81,82].

6. Conclusion

The goal of IMGT/PhyloGene is to automate the phylogenetic analysis of IG and TR genes using Web components and on-line graphical visualisation tools. IMGT/PhyloGene provides fast and relatively accurate reconstructions of phylogenetic trees, and also

provides estimations of synonymous and non-synonymous substitution rates. It has also been designed to be as user-friendly as possible, and does not require the user to possess deep knowledge about phylogenetic analysis. It also does not necessitate to install third-party alignment, reconstruction and vizualisation softwares on the user's computer. IMGT/PhyloGene makes it possible to conduct large scale evolutionary analysis on several hundreds IG and TR genes on the IMGT expertised and standardized data. It also provides a good illustration of the benefits of standardized data, in that they enable sequence comparisons, and constitute an important step before computer based analysis.

Acknowledgements

We are grateful to Véronique Giudicelli, Céline Protat, Denys Chaume and Olivier Gascuel for helpful discussions. Olivier Elemento is supported by a 'Genome' grant from the Ministère de la Recherche. IMGT is funded by the European Union's fifth PCRDT (QLG2-2000-01287) program, the Centre National de la Recherche Scientifique (CNRS), the Ministère de l'Education Nationale and the Ministère de la Recherche. Subventions have been received from Association pour la Recherche sur le Cancer (ARC) and from the Région Languedoc-Roussillon.

References

[1] Lefranc MP. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 2003;31:307–10.
[2] Lefranc M-P. IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. Dev Comp Immunol 2002; 26:697–705.
[3] Lefranc M-P, Lefranc G. The immunoglobulin FactsBook. London: Academic Press; 2001. pp. 1–458, ISBN: 012441351X.
[4] Lefranc M-P, Lefranc G. The T cell receptor FactsBook. London: Academic Press; 2001. pp. 1–398, ISBN: 0124413528.
[5] Ota T, Nei M. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol Biol Evol 1994;11:469–82.
[6] Sitnikova T, Nei M. Evolution of immunoglobulin kappa chain variable region genes in vertebrates. Mol Biol Evol 1998;15:50–60.
[7] Pilström L. The mysterious immunoglobulin light chain. Dev Comput Immunol 2002;26:207–15.
[8] Su C, Jakobsen I, Gu X, Nei M. Diversity and evolution of T-cell receptor variable region genes in mammals and birds. Immunogenetics 1999;50:301–8.
[9] Richards MH, Nelson JL. The evolution of vertebrate antigen receptors: a phylogenetic approach. Mol Biol Evol 2000;17: 146–55.
[10] Glusman G, Rowen L, Lee I, Boysen C, Roach JC, Smit AF, Wang K, Koop BF, Hood L. Comparative genomics of the human and mouse T cell receptor loci. Immunity 2001;15: 337–49.
[11] Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci USA 1997;94:7799–806.
[12] Sitnikova T, Su C. Coevolution of immunoglobulin heavy-and light-chain variable-region gene families. Mol Biol Evol 1998; 15:617–25.
[13] Lefranc M-P, Chuchana P, Dariavach P, Nguyen C, Huck S, Brockly F, Jordan B, Lefranc G. Molecular mapping of the human T cell receptor gamma (TRG) genes and linkage of the variable and constant regions. Eur J Immunol 1989;19: 989–94.
[14] Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. Mol Biol Evol 1999;16: 1315–38.
[15] Pagel M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc (B) 1994;255:37–45.
[16] Pagel M. Inferring evolutionary processes from phylogenies. Zool Scr 1997;26:331–48.
[17] Thompson JD, Higgins DJ, Gibson TJ. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucl Acids Res 1994;22: 4673–80.
[18] Morgenstern B, Frech K, Dress A, Werner T. DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics 1998;14:290–4.
[19] Arden B, Clark SP, Kabelitz D, Mak TW. Human T-cell receptor variable gene segment families. Immunogenetics 1995;42:455–500.
[20] Su C, Nei M. Evolutionary dynamics of the T-cell receptor VB gene family as inferred from the human and mouse genomic sequences. Mol Biol Evol 2001;18:503–13.
[21] Swofford DL. PAUP* Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland, MA, USA: Sinauer Associates; 1999.
[22] Felsenstein J. PHYLIP—PHYLogeny inference package. Cladistics 1989;5:164–6.
[23] Kumar S, Tamura K, Jakobsen IB, Nei M. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 2001; 17:1244–5.
[24] Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Barbié V, Chaume D. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 1999;27:209–12.

[25] Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. Guidelines for human gene nomenclature. Genomics 2002;79:463–70.
[26] Lefranc M-P. Nomenclature of the human T cell receptor genes. Current protocols in immunology. New York: Wiley; 2000. Supplement 40, A.1O.1-A.1O.23.
[27] Lefranc M-P. Nomenclature of the human immunoglobulin genes. Current protocols in immunology. New York: Wiley; 2000. Supplement 40, A.1P.1-A.1P.37.
[28] Lefranc M-P. Unique database numbering system for immunogenetics analysis. Immunol Today 1997;18:509.
[29] Lefranc M-P. The IMGT unique numbering for immunoglobulins, T cell receptors and Ig-like domains. The Immunologist 1999;7:132–6.
[30] Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thévenin-Contet V, Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domain. Dev Comp Immunol 2002;27:55–77.
[31] Lefranc M-P. Nomenclature of the human immunoglobulin heavy (IGH) genes. Exp Clin Immunogenet 2001;18:100–16.
[32] Pallarès N, Lefebvre S, Contet V, Matsuda F, Lefranc M-P. The human immunoglobulin heavy variable genes. Exp Clin Immunogenet 1999;16:36–60.
[33] Scaviner D, Barbié V, Ruiz M, Lefranc M-P. Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. Exp Clin Immunogenet 1999;16:234–40.
[34] Barbié V, Lefranc M-P. The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. Exp Clin Immunogenet 1998;15:171–83.
[35] Lefranc M-P. Nomenclature of the human immunoglobulin kappa (IGK) genes. Exp Clin Immunogenet 2001;18:161–74.
[36] Pallarès N, Frippiat J-P, Giudicelli V, Lefranc M-P. The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. Exp Clin Immunogenet 1998;15:8–18.
[37] Lefranc M-P. Nomenclature of the human immunoglobulin lambda (IGL) genes. Exp Clin Immunogenet 2001;18:242–54.
[38] Scaviner D, Lefranc M-P. The human T cell receptor alpha variable (TRAV) genes. Exp Clin Immunogenet 2000;17:83–96.
[39] Folch G, Scaviner D, Contet V, Lefranc M-P. Protein displays of the human T cell receptor alpha, beta, gamma and delta variable and joining regions. Exp Clin Immunogenet 2000;17:205–15.
[40] Folch G, Lefranc M-P. The human T cell receptor beta variable (TRBV) genes. Exp Clin Immunogenet 2000;17:42–54.
[41] Martinez-Jean C, Folch G, Lefranc M-P. Nomenclature and overview of the mouse (*M. musculus* and *Mus* sp.) immunoglobulin kappa (IGK) genes. Exp Clin Immunogenet 2001;18:255–79.
[42] Bosc N, Lefranc M-P. The mouse (*M. musculus*) T cell receptor beta variable (TRBV), diversity (TRBD) and joining (TRBJ) genes. Exp Clin Immunogenet 2000;17:216–28.
[43] Bosc N, Contet V, Lefranc M-P. The mouse (*M. musculus*) T cell receptor delta variable (TRDV), diversity (TRDD) and joining (TRDJ) genes. Exp Clin Immunogenet 2001;18:51–8.
[44] Lombard V, Camon EB, Parkinson HE, Hingamp P, Stoesser G, Redaschi N. EMBL-Align: a new public nucleotide and amino acid multiple sequence alignment database. Bioinformatics 2002;18:763–4.
[45] Lefranc M-P, Giudicelli V, Busin C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Chaume D. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 1998;26:297–303.
[46] Lefranc M-P. IMGT databases, web resources and tools for immunoglobulin and T cell receptor sequence analysis, *http://imgt.cines.fr*. Leukemia 2003;17:260–6.
[47] Lefranc M-P. IMGT, the international ImMunoGeneTics database. In: Lo B, editor. Antibody engineering protocols, Second edition. Methods in molecular biology, 51. Totowa, NJ, USA: Humana Press; 2003. http://imgt.cines.fr, in press.
[48] Swofford DL, Olsen PJ, Waddell PJ, Hillis DM. Phylogenetic inference. In: Molecular Systematics. Sunderland, MA, USA: Sinauer Associates; 1996. pp.407–514.
[49] Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 1978;27:401–10.
[50] Felsenstein J. Maximum-likelihood and minimum-steps for estimating evolutionary trees from data on discrete characters. Syst Zool 1973;22:240–9.
[51] Huelsenbeck JP. Performance of phylogenetic methods in simulation. Syst Biol 1995;44:17–48.
[52] Jukes TH, Cantor CR. Evolution of protein molecules. In: Mammalian protein metabolism. London, UK: Academic Press; 1969. pp. 21–132.
[53] Kimura M. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980;16:111–20.
[54] Kishino H, Hasegawa MM. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in *Hominoidea*. J Mol Evol 1989;29:170–9.
[55] Felsenstein J, Churchill GA. A hidden markov model approach to variation among sites in rate of evolution. Mol Biol Evol 1996;13:93–104.
[56] Golding GB. Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol Biol Evol 1983;1:125–42.
[57] Yang Z. Phylogenetic analysis by maximum likelihood (PAML). Version 3.0 (http://abacus.gene.ucl.ac.uk/software/paml.html). University College London, London, 2000.
[58] Gojobori T, Nei M. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 1986;3:418–26.
[59] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–25.
[60] Nei M, Kumar S, Takahashi K. The optimization principle in phylogenetic analysis tends to give incorrect topologies when

the number of nucleotides or amino acids used is small. Proc Natl Acad Sci USA 1998;95:12390–7.
[61] Hunkapiller T, Hood L. Diversity of the immunoglobulin gene superfamily. Adv Immunol 1989;44:1–63.
[62] Perrière G, Gouy M. WWW-Query: An on-line retrieval system for biological sequence banks. Biochimie 1996;78: 364–9.
[63] Gilbert DG. PHYLODENDRON: a phylogenetic tree drawing application. Indiana University; 1996.
[64] Hsu E, Steiner LA. Primary structure of Ig through evolution. Curr Opin Struct Biol 1992;2:422–31.
[65] Widholm H, Lundbäck A-S, Daggfeldt A, Magnadottir B, Warr GW, Pilström L. Light chain variable region diversity in Atlantic cod (*Gadus morhua* L.). Dev Comp Immunol 1999; 23:231–40.
[66] Ruiz M, Lefranc M-P. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics DOI 10.1007/s00251-001-0408-6. Immunogenetics 2002;53:857–83.
[67] Swofford DL, Maddison WP. Reconstructing ancestral character states under Wagner parsimony. Math Biosci 1987;87:199–229.
[68] Hassanin A, Golub R, Lewis SM, Lewis GE. Evolution of the recombination signal in the Ig heavy chain variable region locus of mammals. Proc Natl Acad Sci USA 2000;97: 11415–20.
[69] DuBois P. MySQL. Indianapolis, USA: New Riders; 1999.
[70] Kernighan BW, Ritchie DM. The C programming language. Upper Saddle River, USA: Prentice Hall; 1998.
[71] Schwartz RL, Phoenix T. Learning Perl. Cambridge, USA: O'Reilly and Associates; 2001.
[72] Giudicelli V, Chaume D, Bodmer J, Müller W, Busin C, Marsh S, Bontrop R, Lemaitre M, Malik AM, Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 1997;25:206–11.
[73] Giudicelli V, Lefranc M-P. Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 1999;15:1047–54.
[74] Ruiz M, Giudicelli V, Ginestoux C, Stoehr P, Robinson J, Bodmer J, Marsh SG, Bontrop R, Lemaitre M, Lefranc G, Chaume D, Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 2000;28:219–21.
[75] Jones PT, Dear PH, Foote J, Neuberger MS, Winter G. Replacing the complementarity-determining regions in a human antibody with those from a mouse. Nature 1986;321: 522–5.
[76] Singer II, Kawka DW, DeMartino JA, Daugherty BL, Elliston KO, Alves K, Bush BL, Cameron PM, Cuca GC, Davies P. Optimal humanization of 1B4, an anti-CD18 murine monoclonal antibody, is achieved by correct choice of human V-region framework sequences. J Immunol 1993;150:2844–57.
[77] Poul M-P, Ticchioni M, Bernard A, Lefranc M-P. Inhibition of T cell activation with a humanized anti-β1 integrin chain mAb. Mol Immunol 1995;32:101–16.
[78] Rosok MJ, Yelton DE, Harris LJ, Bajorath J, Hellstrom KE, Hellstrom I, Cruz GA, Kristensson K, Lin H, Huse WD, Glaser SM. A combinatorial library strategy for the rapid humanization of anticarcinoma BR96 Fab. J Biol Chem 1996;37: 22611–8.
[79] Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985;39:783–91.
[80] Sitnikova T, Rzhetsky A, Nei M. Interior-branch and bootstrap tests of phylogenetic trees. Mol Biol Evol 1995; 12:319–33.
[81] Elemento O, Gascuel O, Lefranc M-P. Reconstructing the duplication history of tandemly repeated genes. Mol Biol Evol 2002;19:278–88.
[82] Elemento O, Gascuel O. A fast and accurate distance-based algorithm to reconstruct tandem duplication trees. Bioinformatics. Proceedings of European Conference on Computational Biology (ECCB2002). Bioinformatics 2002;18:s92–s99.
[83] Bosc N, Lefranc M-P. The mouse (*Mus musculus*) T cell receptor alpha (TRA) and delta (TRD) variable genes. Dev Comp Immunol 2003;27:465–97.