Conservation of key members in the course of the evolution of the insulin signaling pathway

Janet Piñero González, Olimpia Carrillo Farnés, Ana Tereza R. Vasconcelos, Abel González Pérez

Abstract

Our understanding of the evolution of the insulin signaling pathway (ISP) is still incomplete. One intriguing unanswered question is the explanation of the emergence of the glucostatic role of insulin in mammals. To find out whether this is due to the development of new sets of signaling transduction elements in these organisms, or to the establishment of new interactions between pre-existing proteins, we rebuilt putative orthologous ISPs in 17 eukaryotic organisms. Then, we computed the conservation of orthologous ISPs at different levels, from sequence similarity of orthologous proteins to co-evolution of interacting domains. We found that the emergence of glucostatic role in mammals can neither be explained by the development of new sets of signaling elements, nor by the establishment of new interactions between pre-existing proteins. The comparison of orthologous IRS molecules indicates that only in mammals have they acquired their complete functionality as efficient recruiters of effector sub-pathways.

Keywords: Insulin signaling pathway; IRS; Insulin pathway evolution; Recruiting molecules; Comparative genomics

1. Introduction

The availability of genomic data for a growing number of organisms has led to an increasing number of studies that assess network conservation across different species. Typically, such studies use data on molecular interactions obtained in biochemical and genetic research to extrapolate to other species for which such data is not available with a focus on the evolution of transcriptional regulatory networks and metabolic pathways (Dandekar et al., 1999; Lozada-Chávez et al., 2006).

Likewise, bioinformatics approaches may also be taken to assess the conservation of signal transduction pathways across species, a study that might shed light on the evolution of some well-known signaling pathways, such as the insulin/insulin-like growth factor (IGF) signal transduction pathways (Cohen, 2006). Additionally, reports on the existence of insulin-like peptides and insulin receptor-like molecules in non-mammalian vertebrates and in invertebrates have appeared in scientific literature since the 1970s (Kelley, 1993; Seecof and Dewhurst, 1974; Nagasawa et al., 1984). Biochemical and genetic studies of putative insulin-like signaling pathways in both groups of organisms have focused almost entirely on identifying insulin-like peptides and insulin receptor-like molecules, and the biological effects triggered by the binding of the two (Riehle et al., 2006; Floyd et al., 1999; Lagueux et al., 1990; Satake et al., 1997).
These studies have established that in invertebrates, insulin-like molecules are responsible for stimulating cellular processes that resemble the biological functions of mammalian IGF molecules, such as cell division, protein synthesis, and growth (Chan and Steiner, 2000; Claeyts et al., 2002; Ebberink et al., 1989).

On the other hand, several bioinformatics approaches have been taken to identify putative insulin-like peptides coding genes in several completely sequenced eukaryotic genomes (Duret et al., 1998). In addition, at least one attempt of using bioinformatics tools has been made to identify putative insulin receptor substrate (IRS) molecules in *Caenorhabditis elegans* (CEL) (Wolkow et al., 2002). Moreover, a recent study used the mirror-tree approach to protein–protein interactions assessment (Pazos and Valencia, 2001) to evaluate the probability of the occurrence of the human known cross-reactivity between the IGFs and the insulin receptor in other mammals (Dou et al., 2006).

Several key questions concerning the evolution of the insulin signaling pathway (ISP) remain unanswered. For instance: (1) is the emergence of the glucostatic role of insulin in mammals linked to the development of new sets of signaling transduction elements in this group of organisms, or is it linked to the establishment of new interactions between pre-existing proteins? (2) How does the evolution of the three members that are unique to the pathway (insulin, the insulin receptor, and IRS molecules) has influenced the evolution of the ISP as a whole?

Herein, we address these two questions under the assumption that the differences observed between the biological roles of mammalian insulins and invertebrate insulin-like peptides may be explained through differences either in the composition of orthologous ISPs, or in the pattern of interactions between the elements that form them. We used the human ISP annotated in the KEGG database (Kanehisa et al., 2006) as a template to rebuild putative orthologous ISPs in 17 eukaryotic organisms whose genomes are completely sequenced. These 17 organisms include 7 mammals, 5 non-mammal vertebrates, and 5 invertebrates. Then, we computed the conservation of orthologous ISPs at different levels: domain composition of orthologous proteins; sequence and sequence signatures of orthologous proteins; and co-evolution between interacting pairs of orthologous proteins. We clustered the members of the human ISP (and analogously the members of orthologous ISPs) into five effector pathways that relate to the regulation of different biological processes. Finally, we used these various measurements of conservation of orthologous ISPs to assess whether the five effector pathways within the ISPs are equally conserved across the 17 eukaryotic organisms studied.

Contradicting our initial assumption, we found that all effector pathways are roughly equally conserved. Therefore, the emergence of the glucostatic role of insulin in mammals can neither be explained by the development of new sets of signaling transduction elements, nor by the establishment of new interactions between pre-existing proteins. Nevertheless, we found that as a rule, putative IRS1 orthologs are less conserved among invertebrates than other members of the ISPs, which suggests that only in mammals have IRS molecules acquired their complete functionality as efficient recruiters of effector pathways. This idea is reinforced by highly conserved redundant interactions between orthologous insulin receptor and several effector elements (i.e., bypassing IRS molecules).

### 2. Methods

Amino acid sequences of the proteins that form the human ISP were retrieved from the KEGG database (http://www.genome.jp/kegg; Kanehisa et al., 2006). Other data of these proteins were manually extracted from the Uniprot database (http://www.expasy.org/uniprot; Consortium, 2007).


#### 2.1. Identifying Orthologous Proteins

We blasted (BLASTP) the amino acid sequences of the proteins that form the human ISP according to KEGG against the amino acid sequences of the ORFs of the 17 genomes under study. BLAST (Altschul et al., 1997) searches were performed automatically, and only subject sequences whose alignments returned e-values smaller than $10^{-3}$ with sequence coverage of at least 60% were retrieved. The best hit of the BLAST was considered to be the true ortholog. The human genome was included in the search to obtain the sequences corresponding to translated genes, thus eliminating the noise that might have emerged from post-translational modifications (such as proteolytic processing) in comparative analysis. Moreover, the need to identify the human ORFs that corresponded to the polypeptide sequences reported in KEGG as members of the human ISP rendered unsuitable the Best Bi-directional BLAST Hits (BBH) approach in this case.

To search for insulin and IRS orthologs, we extracted the Hidden Markov’s Models (HMMs) corresponding to the insulin domain (for insulin) and the IRS1 and PH domains (IRS1). Then, we aligned these HMMs to the amino acid sequences of the ORFs of the 17 genomes under study with the HMmer package (version 2.3.2: Eddy, 1998), and retrieved all hits with e-values under $10^{-3}$.

To select the putative true ortholog of the human insulin among the several candidates generated by the HMmer, we first built a multiple alignment of mammalian insulins (selected among all HMmer hits as the highest scoring sequence of each organism in the first alignment of the CLUSTALW program). Then, we used this alignment as a profile to guide the alignment of the candidate insulins from non-mammalian and invertebrate insulins, by also selecting the highest scoring candidate for each organism in the alignment against the previously constructed profile. In the case of IRS1 candidate orthologs, we selected the amino acid sequence that simultaneously produced the alignment with the lowest e-value to the IRS1 HMM and a significant alignment (e-value lower than $10^{-3}$) to the PH HMM. We solely proceeded with the putative orthologs identified using the IRS1 HMM, because searches with the IRS2 HMM produced poorer results.

#### 2.2. Assessment of the Conservation of Domain Composition in Orthologous Proteins

We aligned all the domains in the PFam database against all the proteins identified to be part of the 17 orthologous ISPs. If the alignments of two different domains overlapped in more than 50% of the length of the smaller domain, only the alignment with the lower e-value was retrieved. As in the previous HMM searches, only alignments with e-values lower than $10^{-3}$ were considered significant.
sequences of orthologous proteins were extracted from the database and was available, the whole protein sequence was retrieved from the database. If there was information for either (or both) members of the pair regarding the domain that participates in the interaction across the 17 organisms under study. If there was information for either (or both) members of the pair regarding the domain that participates in the interaction across the 17 organisms under study. If there was information for either (or both) members of the pair regarding the domain that participates in the interaction across the 17 organisms under study. If there was information for either (or both) members of the pair regarding the domain that participates in the interaction across the 17 organisms under study.

2.3. Assessment of Sequence Conservation

We constructed multiple alignments of the sequences of all groups of orthologous proteins within the ISPs using the CLUSTALW software. Then, we computed the number of perfectly conserved positions (i.e., appearance of the same amino acid residue in every sequence within the alignment) in the alignment of the sequences of the 17 orthologs. Also, we computed the number of perfectly conserved positions in the alignment of the sequences of placental mammalian orthologs (six organisms). Finally, we assessed the conservation of amino acid sequence across the orthologs of an ISP member as the ratio between the two previously computed figures (C).

2.4. Assessment of Sequence Signatures Conservation

We manually searched the literature to find sequence signatures of key members of the human ISP and assessed their conservation in orthologous proteins. Table 1 summarizes the protein names, their sequence signatures, and the references from which they were extracted.

### Table 1

<table>
<thead>
<tr>
<th>Protein</th>
<th>Signature</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulin</td>
<td>CCXXXCCXXXXXXCC</td>
<td>Lu et al. (2005)</td>
</tr>
<tr>
<td>Insulin receptor</td>
<td>NPXY</td>
<td>Virkamäki et al. (1999) and Eck et al. (1996)</td>
</tr>
<tr>
<td>IRS (general)</td>
<td>YMXM</td>
<td>White (2002)</td>
</tr>
<tr>
<td>IRS1</td>
<td>GYMPMS, DYMPMS, GYMMMS</td>
<td>White (2002)</td>
</tr>
<tr>
<td>IRS2</td>
<td>GYMPMTP, DYMPMSP</td>
<td>White (2002)</td>
</tr>
</tbody>
</table>

X: any amino acid residue.

2.5. Computing the Conservation of Protein–Protein Interactions

We used the mirror-tree approach (Pazos and Valencia, 2001) to assess the co-evolution of protein pairs, as a measure of the conservation of an interaction across the 17 organisms under study. If there was information for either (or both) member of the pair regarding the domain that participates in the interaction, we retrieved only the stretch corresponding to such domain from the sequence of the protein. If no information on the domains involved in the interaction was available, the whole protein sequence was retrieved from the database. Then, sequences of orthologous proteins were extracted from the database and phylogenetic trees were estimated from CLUSTALW multiple alignments. The co-evolution of trees was assessed through Pearson correlation coefficients:

$$r = \frac{\sum_{i=0}^{16} \sum_{j=0}^{16} (X_{ij} - \bar{X}_i)(Y_{ij} - \bar{Y}_j)}{\sqrt{\sum_{i=0}^{16} \sum_{j=0}^{16} (X_{ij} - \bar{X}_i)^2} \sqrt{\sum_{i=0}^{16} \sum_{j=0}^{16} (Y_{ij} - \bar{Y}_j)^2}},$$

where $X_{ij}$ represents the score of the pairwise alignment of the sequences of the orthologs $i$ and $j$ of the first member of the interacting pair; $Y_{ij}$ represents the same for the second member of the pair; $X_0$ and $Y_0$ represent the mean of all the pairwise scores of the orthologs of the first and second members of the interacting pair, respectively; $n$ is the number of orthologous sequences in each alignment. For pairs whose member proteins had more than one isoform/subunit, more than one pair of matrices was formed, in a combinatorial manner. To assess the statistical significance of these Pearson correlation coefficients, we extracted a random set of 266 human proteins and identified their orthologs in the other 16 organisms as above. Then, pairs of proteins were randomly formed and the correlation coefficients between their alignment matrices were calculated as explained before. We computed the mean of these random correlation coefficients. This process was iterated 1000 times. The mean and standard deviation of the mean correlation coefficient calculated in each iteration were computed. The value of the mean plus two standard deviations of these random values (which we called $r_{ext}$) was applied as a threshold to identify significant correlation coefficients among those calculated for true interacting pairs of proteins within the ISPs.

3. Results and Discussion

The KEGG map (Kanehisa et al., 2006) of the human ISP contains 62 proteins: 17 are protein kinases; 3 are protein phosphatases; 12 are enzymes/carriers that take part directly in metabolic pathways; 3 others are transcription factors; and the remaining are adapters, nucleotide exchange proteins or play other roles. Twenty-four of them present more than one isoform according to KEGG files. In addition, there are five multi-subunit proteins presenting various stoichiometries. Table S1 summarizes important data of these proteins, including the KEGG abbreviations used in this paper.

The activated insulin receptor and/or IRS molecules stimulate five effector pathways according to the biochemical function they regulate (Nelson and Cox, 2004). These pathways are (1) the GLUT4 pathway, formed by eleven proteins that establish 23 interactions (if all isoforms are considered), is involved in the regulation of glucose uptake by muscle and adipose tissue, (2) the phosphoinositide-3-kinase (PI3K) pathway, formed by 14 proteins involved in 59 interactions which participate mainly in the transcriptional regulation of key enzymes of the glycolytic, gluconeogenic and fatty acid biosynthesis pathways, as well as the complex regulation of glycogen synthesis/catabolism, (3) a derivation of the PI3K pathway through the stimulation of the mTOR/Raptor complex formed by 11 proteins establishing 16 interactions involved in the stimulation of protein synthesis, (4) the mitogen activated protein kinase (MAPK) pathway, formed by thirteen proteins involved in 23 interactions takes part in the stimulation of protein and DNA synthesis, (5) there are five proteins that establish 16 interactions involved in the negative regulation of the insulin receptor and IRS which are related to the termination of the insulin signal.

3.1. Orthologous ISPs

There are 137 polypeptidic sequences in the KEGG human ISP map, including isoforms and subunits. These sequences were used to blast translated gene sequences of 17 eukaryotic organisms (obtained from the ENSEMBL database (Hubbard et al., 2007)) in an attempt to rebuild their ISPs (see Section 2 for details). Table 2 presents the names of the 17 organisms included in this study, the abbreviations used to refer to them throughout this paper, and a summary of the results obtained in the reconstruction of their ISPs. Columns 3, 4, 5, and 6 of Table 2 contain the mean, standard deviation, maximum and minimum BLAST e-values of the set of proteins recovered in each organism. (The reported mean and standard deviation correspond to the minus decimal logarithm of the e-values.) Figures S1A and S1B illus-
Table 2
Summary of the results of blasting the KEGG ISP proteins to 17 eukaryotic genomes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Abbreviated name</th>
<th>Mean (−log10 e-value)</th>
<th>S.D. (−log10 e-value)</th>
<th>Maximum e-value</th>
<th>Minimum e-value</th>
<th>Proteins not detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anopheles gambiae</td>
<td>ANG</td>
<td>105.2</td>
<td>57.4</td>
<td>8.0 × 10^-09</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>Apis mellifera</td>
<td>APM</td>
<td>102.2</td>
<td>58.2</td>
<td>2.0 × 10^-07</td>
<td>0</td>
<td>53</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>BTA</td>
<td>148.7</td>
<td>45.3</td>
<td>3.0 × 10^-08</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>CEL</td>
<td>94.8</td>
<td>57.2</td>
<td>2.0 × 10^-06</td>
<td>0</td>
<td>62</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>CAF</td>
<td>156.2</td>
<td>41.7</td>
<td>2.0 × 10^-15</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Ciona intestinalis</td>
<td>CIN</td>
<td>112.9</td>
<td>56.3</td>
<td>2.0 × 10^-08</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>DRE</td>
<td>138.9</td>
<td>50.9</td>
<td>7.0 × 10^-09</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>DME</td>
<td>106.2</td>
<td>58.3</td>
<td>4.0 × 10^-07</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td>Fugu rubripes</td>
<td>FRU</td>
<td>141.7</td>
<td>49.7</td>
<td>6.0 × 10^-09</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>GAL</td>
<td>144.6</td>
<td>48.1</td>
<td>4.0 × 10^-10</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>HSA</td>
<td>158.7</td>
<td>42.2</td>
<td>8.0 × 10^-17</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Monodelphis domestica</td>
<td>MOD</td>
<td>153.0</td>
<td>42.7</td>
<td>4.0 × 10^-20</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>MUS</td>
<td>158.2</td>
<td>34.6</td>
<td>3.0 × 10^-14</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Pan troglodites</td>
<td>PTR</td>
<td>149.7</td>
<td>46.2</td>
<td>1.0 × 10^-16</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>RNO</td>
<td>155.9</td>
<td>43.2</td>
<td>2.0 × 10^-14</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Tetraodon nigroviridis</td>
<td>TEN</td>
<td>137.7</td>
<td>49.3</td>
<td>2.0 × 10^-28</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>Xenopus tropicalis</td>
<td>XET</td>
<td>145.5</td>
<td>46.5</td>
<td>2.0 × 10^-11</td>
<td>0</td>
<td>23</td>
</tr>
</tbody>
</table>

The last column of the table presents the number of polypeptide sequences that the BLAST searches failed to find. Thirteen orthologs were not found within the human genome, while only approximately 20% of invertebrates' orthologs have been experimentally identified as significant insulin-like peptides. The proteins annotated in KEGG and Uniprot (Consortium, 2007), the proteins from the human IRS1 (Chen and Steiner, 2000) and insulin-like pathways, which are hard to predict from the translated genes, are among those missed by the query ISP proteins. Revealing the discrepancies between the invertebrate and human proteomes, the insulin-like peptides derived from gene duplication events. The proteins from the human IRS1 (Chen and Steiner, 2000) and insulin-like pathways, which are hard to predict from the translated genes, are among those missed by the query ISP proteins. Revealing the discrepancies between the invertebrate and human proteomes, the insulin-like peptides derived from gene duplication events.
Table 3

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of significant HMM hits</th>
<th>e-value</th>
<th>Maximum e-value</th>
<th>Minimum e-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANG</td>
<td>3</td>
<td>3.0 × 10^{-2}</td>
<td>8.7 × 10^{-1}</td>
<td>1.7 × 10^{-2}</td>
</tr>
<tr>
<td>BTA</td>
<td>2</td>
<td>6.4 × 10^{-3}</td>
<td>8.7 × 10^{-1}</td>
<td>2.1 × 10^{-2}</td>
</tr>
<tr>
<td>CEL</td>
<td>6</td>
<td>4.6 × 10^{-2}</td>
<td>7.4 × 10^{-3}</td>
<td>2.1 × 10^{-2}</td>
</tr>
<tr>
<td>CIN</td>
<td>7</td>
<td>2.9 × 10^{-1}</td>
<td>1.4 × 10^{-1}</td>
<td>2.1 × 10^{-2}</td>
</tr>
<tr>
<td>DAR</td>
<td>11</td>
<td>1.2 × 10^{-2}</td>
<td>1.1 × 10^{-2}</td>
<td>2.1 × 10^{-2}</td>
</tr>
<tr>
<td>DME</td>
<td>8</td>
<td>5.1 × 10^{-2}</td>
<td>1.4 × 10^{-1}</td>
<td>2.1 × 10^{-2}</td>
</tr>
<tr>
<td>GAL</td>
<td>5</td>
<td>2.1 × 10^{-2}</td>
<td>1.4 × 10^{-1}</td>
<td>2.1 × 10^{-2}</td>
</tr>
<tr>
<td>HIS</td>
<td>3</td>
<td>1.3 × 10^{-2}</td>
<td>1.4 × 10^{-1}</td>
<td>2.1 × 10^{-2}</td>
</tr>
<tr>
<td>MOD</td>
<td>5</td>
<td>4.9 × 10^{-2}</td>
<td>1.4 × 10^{-1}</td>
<td>2.1 × 10^{-2}</td>
</tr>
<tr>
<td>MUS</td>
<td>6</td>
<td>3.1 × 10^{-2}</td>
<td>1.4 × 10^{-1}</td>
<td>2.1 × 10^{-2}</td>
</tr>
<tr>
<td>PTR</td>
<td>13</td>
<td>4.4 × 10^{-2}</td>
<td>1.4 × 10^{-1}</td>
<td>2.1 × 10^{-2}</td>
</tr>
<tr>
<td>RNO</td>
<td>11</td>
<td>3.1 × 10^{-2}</td>
<td>1.4 × 10^{-1}</td>
<td>2.1 × 10^{-2}</td>
</tr>
</tbody>
</table>

As presented in Table 3, three out of the four organisms with the higher number of hits for the IRS1 HMM are fish. This is probably due to a genome duplication event occurred in this group (Taylor et al., 2003) and agrees with previous reports (Dou et al., 2006).

It is important to point out that the CEL ortholog found by our methodology differs from the protein detected by Pertseva and Shpakov (2002) in the aforementioned work (ist1). The ist1 and the IRS1 HMM produce a very low score alignment (−8.6), being unlikely for this protein to interact with phosphorytrosine residues in the insulin receptor.

3.2. Conservation of Orthologous ISPs

We used the scores and e-values returned by the HMMer package (Eddy, 1998) to explore the conservation of 93 domains in the polypeptide sequences that form the 17 orthologous ISPs previously reconstructed (see Section 2 for details). Figure S2A presents the distribution of e-values (as minus decimal logarithm) resulting from the alignments of the 93 corresponding HMMs to the proteins of the 17 orthologous ISPs. Some domains are found in more than one protein, as may be appreciated by vertical stretches of multicolor dots (each color corresponding to one organism). Such is the case of the protein tyrosine kinase and the general protein kinase domains that appear in multiple members of the ISPs. Despite their wide distribution, these two domains are amongst the 31 better conserved domains in the orthologous ISPs, presented in Figure S2B. This property is shared by other domains associated with the functionality of the signaling pathway, such as the protein phosphatase domain and may be explained by the fact that parts of the ISP (such as the MAPK pathway) may have appeared early in the evolution of eukaryotes (Widmann et al., 1999).
and computed the ratio of the number of conserved residues in these alignments to the number of conserved residues in the alignments of the sequences of mammalian orthologs. We called this ratio \( C \). We excluded MOD (a marsupial) from mammalian sequences alignments due to the divergence of its protein sequences with respect to the other mammalian orthologs (Hulsen et al., 2006). Figure S3 shows the frequency histogram of \( C \) for the ISP proteins. Two thirds of ISP proteins possess \( C \) ratios between 0 and 0.2, i.e., the conservation of the sequence of their orthologs is, as a rule, less than 20% the conservation of mammalian orthologs. Since \( C \) is a ratio of numbers of residues, these values mean that only relatively short sequence stretches have been conserved in the course of the evolution of these proteins. On the other extreme of this distribution, two percent of the ISP orthologous proteins exhibit a degree of conservation that extends over more than 70% of amino acid residues conserved among mammalian orthologs. Again, three enzymes involved in the regulation of glycogen synthesis/catabolism (PHK, PKA, and PP1) display the \( C \) values that correspond to higher sequence conservation.

A quick glance at Fig. 1 reveals that there are no apparent differences in the conservation of orthologous sequences across the ISPs. None of the five abovementioned pathways exhibits uncolored rectangles that would correspond to proteins with orthologous subunits/isoforms found in less than 15 organisms (the only exception is BAD). Moreover, although at first sight the proteins that form the MAPK pathway seem to be better conserved in average than the proteins in the other pathways, no clear trend of higher co-evolution of pairs of proteins could be established (see below).

Given these results, we proceeded to evaluate the co-evolution of pairs of proteins that interact in the human ISP, in an attempt to discern whether differences in insulin action between vertebrates and invertebrates could be explained by a differential conservation of the pattern of protein-protein interactions among the five pathways.

We collected all interactions established between pairs of proteins according to the KEGG map of the human ISP. For proteins with more than one subunit/isoform, more than one pair of possible interacting proteins was formed, in order to evaluate all possibilities of co-evolution between pairs of proteins. We assessed the co-evolution of these pairs of proteins employing the mirror-tree approach (Pazos and Valencia, 2001). Briefly, this strategy is based on the fact that interacting protein pairs are subjected to selective pressure on the conservation of such interaction. In some cases this selective pressure would pro-

![Diagram](image-url)
duce a pattern of co-evolution of the two interacting proteins, mainly because of compensatory mutations. This pattern may be detected by aligning the sequences of a number of orthologs of the two interacting proteins, and computing the distances between the leaves of the phylogenetic trees that correspond to the two alignments.

Here, we started with a number of well-documented experimentally tested interactions between pairs of human proteins that form the ISP, and used the mirror-tree approach (Pazos and Valencia, 2001) to evaluate whether a trace of co-evolution could be detected from the comparison of the alignments of the 17 orthologs of each protein of the interacting pair. Following the same idea, a recent study has used the mirror-tree approach to assess the existence of cross-reactivity between insulin and IGF in a number of mammals (Dou et al., 2006). To assess the statistical significance of the correlation coefficients calculated from the phylogenetic trees of pairs of proteins we compared them to the mean value plus two standard deviations of a set of correlation coefficients calculated from 1000 randomly chosen sets of pairs of human proteins and their orthologs (a value that we called $r_{ext}$).

Figure S4 presents the values of correlation coefficients of phylogenetic trees calculated for each pair of polypeptide sequences (133) formed as explained above. The horizontal black line corresponds to the value of $r_{ext}$. The mean of these randomly calculated correlation coefficients is greater than 0 (0.54), probably because any set of proteins selected from the same genome show a certain degree of co-evolution (Pazos et al., 2005). Fifty-one pairs of the phylogenetic trees computed from the alignments of orthologous sequences exhibit correlation coefficients greater than $r_{ext}$, despite the strength of this threshold value. Nevertheless, Fig. 1 suggests that these 51 pairs of interacting polypeptide sequences are homogeneously distributed over the five pathways.

Table 4 presents a comparison of the mean, standard deviation, maximum and minimum values of the sets of correlation coefficients of pairs of phylogenetic trees calculated for pairs of proteins of the five pathways separately. In this inter-pathway comparison we excluded interactions that are common to two or more pathways (those between insulin and its receptor, or between the receptor and IRS1, for example).

A quick inspection of Figure S4 suggests that the co-evolution of interacting protein pairs (as evaluated through correlation of phylogenetic trees) is roughly homogeneous across the five pathways. A more rigorous measurement, presented in Table 4 as percentage of correlation coefficient values that are greater than $r_{ext}$ in each pathway, reveals differences among pathways. On average 40% of all protein pairs have correlation coefficients higher than $r_{ext}$, whereas 70% of the protein pairs involved in the termination of the insulin signal present correlation coefficient above this value. This suggests that these proteins tend to co-evolve in a stronger manner than all the members of the ISP.

To assess the significance of such trend, we compared the correlation coefficients calculated for the five pathways using the Kruskall–Wallis test. The results of this analysis (Fig. 2) showed that, at $p < 0.05$ only the Regulation pathway and the GLUT4 pathway are significantly different (in terms of the distribution of correlation coefficients), while all other pathway pairs do not present any significant differences. At $p < 0.01$ no significant differences are observed among pathways.

These results indicate that our original hypothesis must be ruled out, because we were unable to identify significant differences in the conservation of the five effector pathways within the ISP. It is likely that these five pathways have developed early in the evolution of eukaryotic organisms – even before the emergence of any ISP – and respond to various stimuli. Some of them, as the MAPK pathway, are present even in Saccharomyces cerevisiae (Barbieri et al., 2003; Conlon, 2001).

### 3.3. IRS Molecules and the Evolution of the ISP

In the last part of the work, we focused on the three proteins that are unique to the ISP, namely insulin, the insulin receptor, and the IRS. We analyzed the conservation of sequence signatures that have been characterized in their mammalian representatives across the 17 orthologs identified here.

The motif CCX$_3$CX$_2$C, which is characteristic to insulin, is conserved across all its orthologs (Fig. 3A). Despite their relatively low sequence conservation ($C = 0.045$), pointed out by
Fig. 3. Conservation of sequence signatures in orthologous ISP proteins. Only relevant portions of sequence alignments are shown. Amino acid residues that form sequence signatures are highlighted in gray. Regular expressions for these sequence signatures and their original sources are presented in Table 1. (A) Insulin, (B) insulin receptor and (C) IRS1.

several previous works (Claeys et al., 2002; Duret et al., 1998; Conlon, 2001; Smit et al., 1998) all insulin orthologs showed approximately the same number of residues, and a clear trend could be observed that favored the conservation of the sequence stretches corresponding to the A and B peptides (Fig. 3A) over the rest of the sequence.

In the case of receptor, all orthologs, except the one of ANG presented the signature NPXY (Fig. 3B), which is recognized by various proteins that interact with the receptor (White et al., 1988). The analysis of domain composition of receptor orthologs (Figure S5) showed that the protein tyrosine kinase domain is better conserved across all organisms, while the other domains are more poorly conserved, and some of them are even absent in some organisms. In the ANG ortholog, for example, all domains except the protein tyrosine kinase are absent, which explains the failure in detecting the sequence signature. The absence of several domains in the RNO ortholog is probably due to an error in the annotation of its ORF. A fragment of the RNO mRNA encoding the insulin receptor carrying 46 more amino acid residues than the protein identified in this study is annotated in the NCBI
The sequences of putative IRS1 orthologs detected in invertebrate organisms are, as a rule, shorter than their vertebrate counterparts (360 amino acid residues on average vs. 1200 residues). They overlap at the amino terminal region of mammalian IRS1, where the IRS and PH domains are located. Indeed, Figure S5 shows that the IRS1 domain was detected in all orthologous proteins (and the PH domain in all except the CEL ortholog). On the other hand, the C-terminal region, known to contain small sequence motifs involved in activating downstream substrates in the ISP (White, 2002) is absent in most invertebrate orthologs (Figure S6). Taken together, these results suggest that no true IRS molecules (in the functionality they exhibit in mammalian ISPs) exist in the invertebrate organisms included in our study. This idea is reinforced by the finding that probably very well conserved interactions (with high correlation coefficient values that indicate strong co-evolution) occur between the insulin receptor and downstream substrates in the ISPs. These interactions are redundant to those mediated by IRS molecules, suggesting that the latter may have emerged at some point of the evolution of the ISP, amplifying and adding versatility to signaling through various effector pathways. The idea that IRS molecules might have appeared in the evolution of the ISPs in certain invertebrate groups had been put forward by Pertseva and Shpakov (2002). Here, we take a step forward and show that proteins with domains of the IRS and PH types (a branch of phosphotyrosine binding domains) already existed in at least some invertebrates. Their evolution could have proceeded by fusion to further domains (with various phosphotyrosine-containing sequence stretches) to conform the complex IRS molecules found in mammals which are involved in the interaction with different types of downstream substrates.

4. Conclusions

All downstream effector pathways of the insulin cascade are roughly equally conserved across vertebrates and invertebrates. This implies that the emergence of the glucostatic role of insulin in mammals can neither be explained by the development of new sets of signaling transduction elements, nor by the establishment of new interactions between pre-existing proteins in the course of the evolutionary process. On the other hand, putative IRS1 orthologs are clearly less conserved among invertebrates with respect to their mammalian counterparts than other members of the ISPs. Processes of gene duplication and fusion may have shaped IRS molecules during evolution, as they acquired their role of efficient and versatile recruiters of downstream effector pathways shown in mammals. Direct interactions between the activated insulin receptor and downstream effector elements may have been predominant in invertebrates.

Acknowledgments

Authors are thankful to Marcia Triunfol (LNCC) for careful revision of the manuscript. AGP is grateful to the Havana database under Accession number XM_512323 (Pruitt et al., 2007).

The Swiss Development and Cooperation Agency for meaningful support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.biosystems.2008.06.003.

References


Krug, R.L., Dewhurst, S., 1974. Insulin is a Drosophila hormone and acts to enhance the differentiation of embryonic *Drosophila* cells. Differentiation 3, 63–70.


