BMC
Structural Biology

A three-dimensional topology of complex I inferred from evolutionary correlations

Philip R Kensche[1,2*], Isabel Duarte[1,2] and Martijn A Huynen[1,2]

Abstract

Background: The quaternary structure of eukaryotic NADH:ubiquinone oxidoreductase (complex I), the largest complex of the oxidative phosphorylation, is still mostly unresolved. Furthermore, it is unknown where transiently bound assembly factors interact with complex I. We therefore asked whether the evolution of complex I contains information about its 3D topology and the binding positions of its assembly factors. We approached these questions by correlating the evolutionary rates of eukaryotic complex I subunits using the mirror-tree method and mapping the results into a 3D representation by multidimensional scaling.

Results: More than 60% of the evolutionary correlation among the conserved seven subunits of the complex I matrix arm can be explained by the physical distance between the subunits. The three-dimensional evolutionary model of the eukaryotic conserved matrix arm has a striking similarity to the matrix arm quaternary structure in the bacterium *Thermus thermophilus* (rmsd=19 Å) and supports the previous finding that in eukaryotes the N-module is turned relative to the Q-module when compared to bacteria. By contrast, the evolutionary rates contained little information about the structure of the membrane arm. A large evolutionary model of 45 subunits and assembly factors allows to predict subunit positions and interactions (rmsd = 52.6 Å). The model supports an interaction of NDUFAF3, C8orf38 and C2orf56 during the assembly of the proximal matrix arm and the membrane arm. The model further suggests a tight relationship between the assembly factor NUBPL and NDUFA2, which both have been linked to iron-sulfur cluster assembly, as well as between NDUFA12 and its paralog, the assembly factor NDUFAF2.

Conclusions: The physical distance between subunits of complex I is a major correlate of the rate of protein evolution in the complex I matrix arm and is sufficient to infer parts of the complex's structure with high accuracy. The resulting evolutionary model predicts the positions of a number of subunits and assembly factors.

Keywords: Eukaryotic complex I, Quaternary topology, Assembly, Mirror-tree method, Co-evolution

Background

NADH:ubiquinone oxidoreductase (complex I) is with about 1000 kDa [1,2] the largest of the five complexes of the oxidative phosphorylation (OXPHOS) and a major contributor to the proton motive force that drives the ATP production by ATP-synthase [3]. Complex I has an L-shape with a hydrophilic matrix arm that protrudes into the cytoplasm in bacteria or the mitochondrial matrix in eukaryotes and a hydrophobic membrane arm.

The canonical "core" of complex I consists of 14 subunits that originate from three pre-existing evolutionary modules [4]. The N-module at the distal end of the matrix arm contains flavin-mononucleotide (FMN) that accepts electrons from a donor, usually NADH. The electrons are transported through a chain of iron-sulfur (FeS) clusters along the matrix arm towards the joint of the two arms at the membrane. This membrane-proximal part of the matrix arm represents the Q-module in which the electrons are transferred to ubiquinone (Q). The energy freed by the electron-transfer is transmitted along the P-module (NADH1-6/4L) that uses the energy to pump protons across the membrane [5-7].

In diverse taxa, the canonical core of complex I has been extended by further subunits. For instance, complex I in

---
* Correspondence: pkensche@cmbi.ru.nl
[1]Center for Molecular and Biomolecular Informatics / Nijmegen Center for Molecular Life Sciences, Radboud University Medical Center, PO Box 9101, Nijmegen, HB 6500, The Netherlands
[2]Netherlands Bioinformatics Centre, Geert Grooteplein 28, Nijmegen, GA 6525, The Netherlands

Thermus thermophilus contains an additional subunit located at the interface of the N- and Q-modules [8] and a recent analysis of complex I in the α-proteobacterium *Paracoccus denitrificans* identified three additional subunits [9]. Eukaryotes obtained complex I with the endosymbiotic uptake of an α-proteobacterium that gave rise to the mitochondria. Following the endosymbiosis, the mitochondrial genome was reduced and the genes encoding matrix arm subunits of complex I were transferred to the nucleus. Additionally, complex I was extended to up to 45 subunits by so-called "accessory" or "supernumerary" subunits [1,10]. This set of permanent subunits is further extended by a number of assembly factors absent from the mature complex [11-19].

Up to now, the structures of the complete complex in the eubacterium *Thermus thermophilus* [5] and the eukaryote *Yarrowia lipolytica* [6] have been published. However, the latter structure is of a too low resolution to allow identification of the positions of the supernumerary subunits. Approximate subunit positions within the eukaryotic complex are hinted at by various types of experiments, mostly from sub-complexes observed by fractionation or as assembly intermediates (e.g. [1,20]). For instance, the application of chaotropic detergents to the bovine complex produces the three sub-complexes Iα, Iλ, and Iγ. Because these sub-complexes are large, they provide only rough information about subunit positions. For instance, Iα represents an extended Iλ sub-complex and the additional subunits could in principle be located anywhere on the surface of the Iλ sub-complex. Only limited data are available from yeast-two-hybrid [12,21], co-immunoprecipitation [14], or cross-linking [22] experiments. The identification of the positions of the assembly factors is hampered by the temporariness of the assembly intermediates and our incomplete understanding of the assembly process.

The increasing number of genome sequences allows making predictions of physical interactions by evolutionary correlation methods, including the co-occurrence of genes or phylogenetic profiling [23-26], the mirror-tree approach [27], and residue correlation [28-31], which have successfully identified new complex I subunits and assembly factors [21,25] (reviewed in [32]) and predicted relations between the five OXPHOS complexes [33]. Of these methods, residue correlation is based on the most direct evidence of physical interaction, namely the compensatory mutations at sites of interacting residues to maintain the structure of a protein or complex. By contrast, the mirror-tree method detects co-evolution more indirectly by correlating sequence similarity matrices between orthologous groups [27]. The similarities between protein sequences depend both on species divergence times and on rates of evolution. By removing the similarity due to the species divergence times [34,35], one obtains

similarities that are more related to evolutionary rates. A high correlation in evolutionary rates between protein families can be evidence of a direct physical relation between proteins. For instance, if there is selection to maintain the interaction of two proteins then disrupting mutations have to be compensated for at the rate that they occur. Therefore, to maintain the interaction, an increased rate of change in one protein needs to be compensated for *at a similar rate* in the other protein. Note that the pairwise correlations in evolutionary rates between proteins can also be due to indirect interactions [36].

Here we ask whether we can use evolutionary rate correlations to predict the three-dimensional (3D) conformation of complex I. To this aim, we analyzed the evolutionary rate correlation between 38 subunits and 7 assembly factors of human complex I using the mirror-tree method [35] and find that subunits that are known to be physically close in complex I tend to show a higher degree of correlation in evolutionary rates than those that are physically distant. In the conserved core of the matrix arm, this correlation is strong enough to construct a 3D model with striking similarity to the bacterial reference structure. In a second evolutionary model that includes the 14 canonical core subunits, the membrane and matrix arms appear as clearly separated groups. Finally, we calculated a third evolutionary model including 38 subunits and seven assembly factors. This last model retains some features of the physical structure, including the separation of the matrix and membrane arms and a proximodistal axis in the matrix arm. We discuss the positions of the seven assembly factors in this model and make specific predictions about the association of some assembly factors with each other and with the permanent subunits.
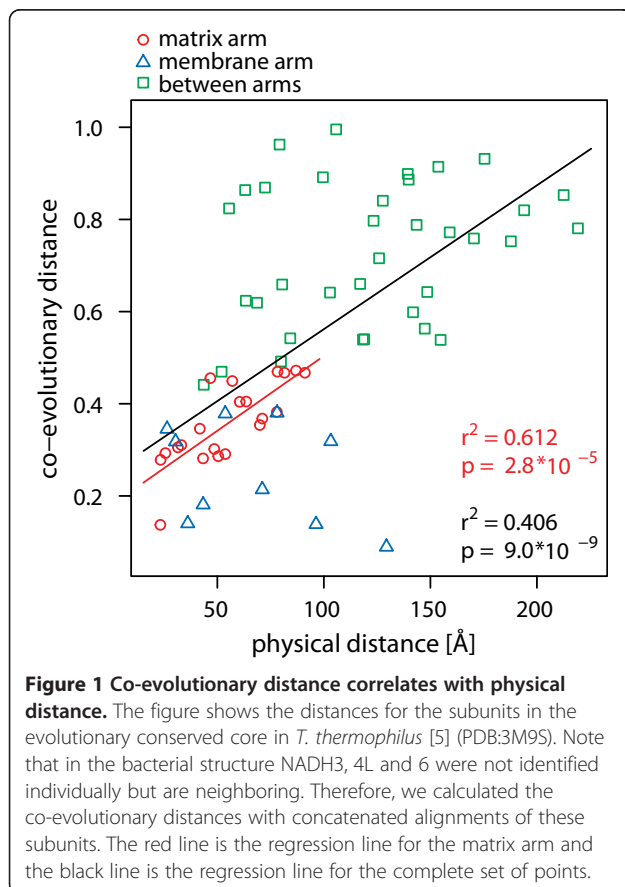
Results
Correlation of physical and co-evolutionary distances in the conserved core

Because our study aims at complex I of human we selected 38 subunits and seven assembly factors of the human complex that have a sufficient number of orthologs for the application of the mirror-tree method (Additional file 1: Table S1). Shortly, we collected homologs of the 45 proteins by querying the nr database. All orthologous sequence sets were aligned and highly variable alignment columns were filtered out using BMGE [37]. Note that NADH3, 4L and 6 could not unambiguously be located in our reference – the structure of the complete complex I of *T. thermophilus* [5] (PDB:3M9S). However, these subunits are known to be direct neighbors and we decided to treat them as a single unit, termed NADH34L6. We calculated maximum-likelihood trees from the alignments [38] and obtained a distance matrix for each protein family from which we removed the common signal of the phylogeny [35]. The phylogeny-corrected matrices were correlated and the resulting
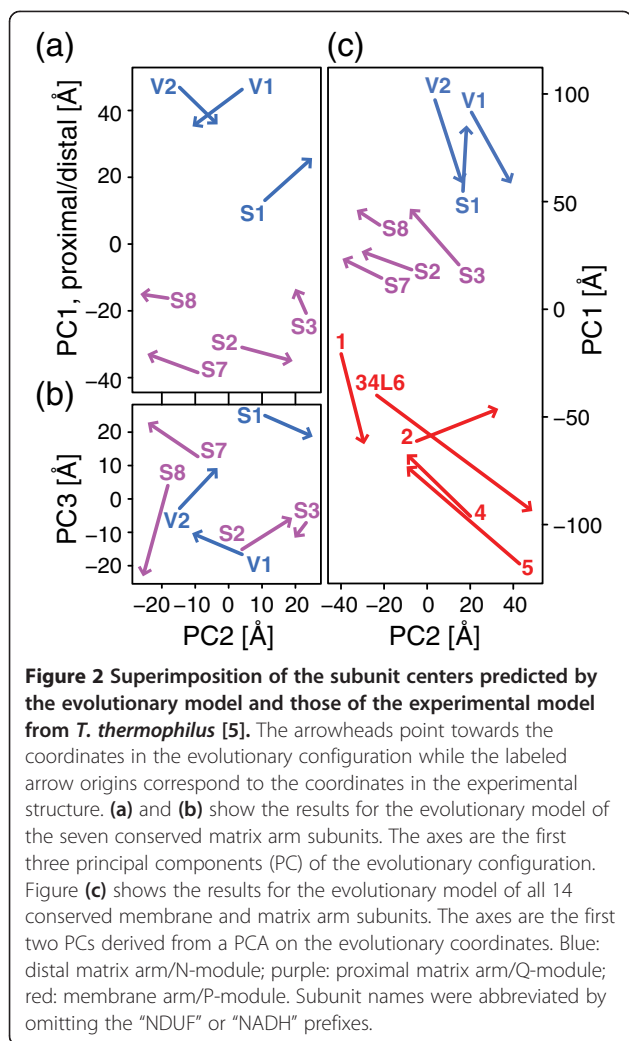
correlation matrix was transformed into the distance matrix (see Methods). For the purpose of this article we call these distances "co-evolutionary" but note that the signal measured by the mirror-tree method is also determined by expression and general functional relatedness [39,40].

First we examined how well the co-evolutionary distances correspond with the distances between the 14 subunits of the conserved core of complex I (Figure 1). We compared the co-evolutionary distances to the distances of the centers of mass of the protein (see Methods). We will refer to these latter distances as "physical distances" and to the arrangement of the subunit centers in three dimensions as the "quaternary topology" of complex I. A direct comparison of the two distance measures reveals that more than 40% of the variation in the co-evolutionary distances can be explained by the physical distances of the subunit ($r^2=0.41$, $p=9\times10^{-9}$, n=66; Figure 1). Within the matrix arm, physical distance explains more than 60% of the variation in co-evolutionary distances (Figure 1; $r^2=0.61$, $p=2.93\cdot10^{-5}$, n=21). Although there is no significant correlation among the membrane arm subunits ($r^2=0.13$, $p=0.23$, n=10), the between-arm co-evolutionary distances clearly are larger than the within-arm co-evolutionary distances, reflecting the physical separation of the two arms.

Figure 1 Co-evolutionary distance correlates with physical distance. The figure shows the distances for the subunits in the evolutionary conserved core in *T. thermophilus* [5] (PDB:3M9S). Note that in the bacterial structure NADH3, 4L and 6 were not identified individually but are neighboring. Therefore, we calculated the co-evolutionary distances with concatenated alignments of these subunits. The red line is the regression line for the matrix arm and the black line is the regression line for the complete set of points.

The high correlation of co-evolutionary and physical distance suggests that it may be possible to obtain an accurate 3D model of the protein complex from the pairwise co-evolutionary distances. We used classical multidimensional scaling (cMDS) to integrate the co-evolutionary distances between the seven matrix arm core subunits into an evolutionary 3D configuration. To ensure that the 3D configuration reliably reflects the co-evolutionary distances, we calculated the P_3-value of the configuration, a cMDS-specific goodness-of-fit measure that is analogous to the fraction of variation explained by the first three eigenvalues in a principal component analysis [41] (see Methods). The P_3-value of the evolutionary configuration is 0.89 and thus close to that of a perfect fit (1.0) and well above the cutoff 0.8 suggested as desirable [41]. Next, we compared the evolutionary 3D model with the bacterial structure (Figure 2a; see Methods). The correlation of the distances in the 3D configuration with the distances in the bacterial structure is lower ($r^2=0.55$) than that of the raw co-evolutionary distances ($r^2=0.61$) but still significant ($p=1.28\cdot10^{-4}$, n=21). The root mean square deviation (rmsd) of the bacterial quaternary topology with the eukaryotic evolutionary configuration is 18.7 Å, which compares well to the about 180 Å length of the matrix arm [8]. We subjected the evolutionary configuration to a principal component analysis and used the principal components thus obtained as new coordinate system for the configuration (Figure 2a/b). The proximodistal axis of the matrix arm core corresponds to the first and largest principal component (Figure 2a) and is thus correctly identified by the evolutionary model as the axis with the largest extent. The positioning of subunits along this axis is almost perfectly recovered by the evolutionary model (correlation bacterial/predicted along the first axis: $r^2=0.95$, $p=2.1\times10^{-4}$, n=7). Also the second axis shows a strong correlation ($r^2=0.62$, $p=3.5\times10^{-2}$) while the correlation along the third axis ($r^2=0.45$) is significant at a level of 9.8% ($p=9.8\times10^{-2}$). The projection of the second and third axes (Figure 2b) shows that among both the proximal Q-module subunits and distal N-module subunits the evolutionary model identifies the correct circular ordering around the proximodistal axis. Furthermore, in the predicted model, which is based on eukaryotic sequences only, the four Q-module subunits (Figure 2b; purple) are twisted relative to the three N-module subunits (Figure 2b; blue) compared to the bacterial structure. Interestingly, a twist in the same direction was observed in a comparison of the matrix arms of the eukaryote *Y. lipolytica* and *T. thermophilus* ([6], personal communication): When looking from the matrix towards the membrane the N-module is turned clockwise relative to the Q-module.

After the prediction of the topology of the seven matrix arm core subunits, we predicted the topology of the complete core of 14 matrix- and membrane arm subunits
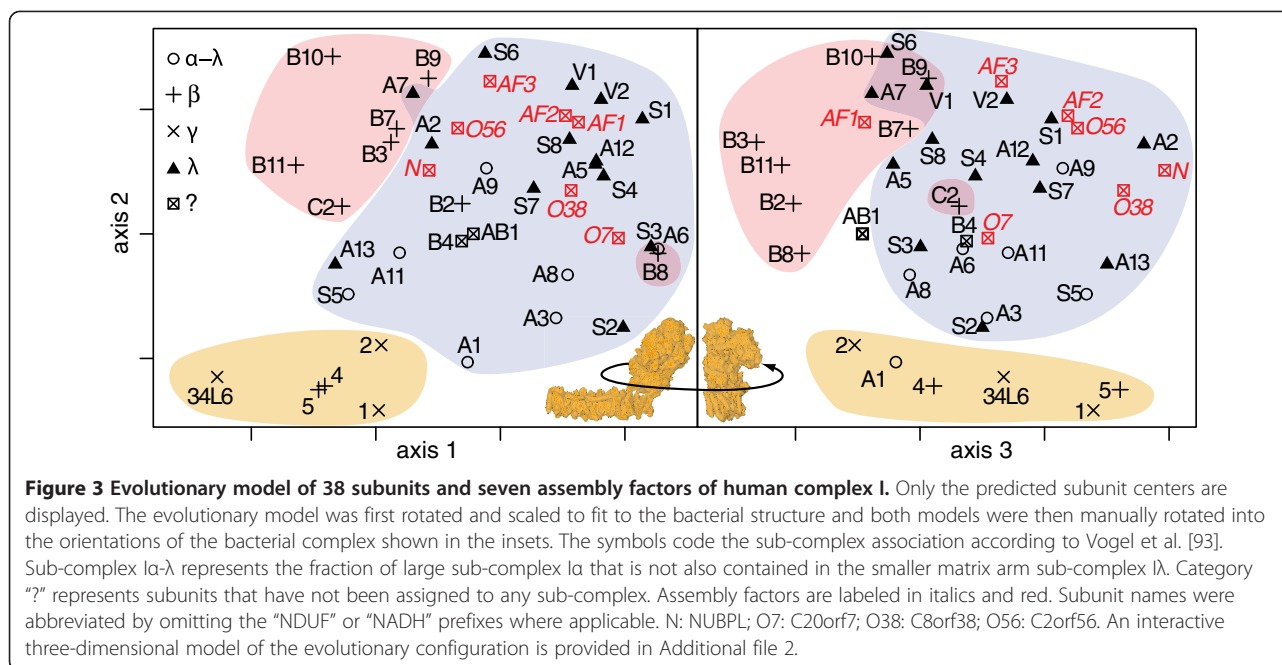
Figure 2 Superimposition of the subunit centers predicted by the evolutionary model and those of the experimental model from *T. thermophilus* [5]. The arrowheads point towards the coordinates in the evolutionary configuration while the labeled arrow origins correspond to the coordinates in the experimental structure. **(a)** and **(b)** show the results for the evolutionary model of the seven conserved matrix arm subunits. The axes are the first three principal components (PC) of the evolutionary configuration. Figure **(c)** shows the results for the evolutionary model of all 14 conserved membrane and matrix arm subunits. The axes are the first two PCs derived from a PCA on the evolutionary coordinates. Blue: distal matrix arm/N-module; purple: proximal matrix arm/Q-module; red: membrane arm/P-module. Subunit names were abbreviated by omitting the "NDUF" or "NADH" prefixes.

that are conserved among bacteria and eukaryotes. Again, the co-evolutionary distances of the subunits can be well embedded in three dimensions (P_3=0.84) and result in a configuration with separate membrane and matrix arms (Figure 2c). The rmsd of this evolutionary model and bacterial structure is 47.6 Å, which corresponds to about 25% of the length of the membrane arm [5]. Also in the complete-core model, the proximodistal axis of the matrix arm is recovered (r^2=0.96, p=9.3×10^{-5}) although the rmsd of the matrix arm core subunits to the bacterial structure is lower (34.0 Å) than that for the configuration that only contained matrix arm subunits (18.7 Å). As expected from the correlation between the pairwise physical distances and their evolutionary correlation, the accuracy of the positioning in the membrane arm is poor (rmsd=61.8 Å). A closer inspection of the evolutionary model suggests that the membrane arm subunits are located as a cluster more sideways of the matrix arm than in the physical structure. In the superimposition of the two structures, this manifests in a tendency of the Q-module away from

the membrane and a slight tendency of the distal N-module towards the membrane.

Eukaryotic complex I and assembly factors

Although the structure of the 14 core subunits in bacteria is known [5] and largely conserved in eukaryotes [6], the arrangement of the mostly eukaryote-specific accessory subunits has not yet been resolved. Furthermore, the transient nature of the interaction of assembly factors in complex I assembly intermediates hampers the identification of their binding sites in the complex. We therefore asked whether the positions of the accessory subunits and assembly factors could be identified from their evolutionary correlation. We extended the evolutionary 3D configuration to include 38 permanent subunits and 7 assembly factors of the human complex. Both the goodness-of-fit of the 3D configuration with the raw co-evolutionary distances and the comparison with the reference structure, indicate the quality of this model. The goodness-of-fit measure P_3, which expresses how well the 3D model represents the pairwise co-evolutionary distances, is lower (P_3=0.43) than for the previous models. However, a comparison of this P_3-value with the distribution of P_3-values of 10^6 permuted symmetric matrices fitted in 3D (P_3^{random}=0.17±0.005) shows that the arrangement of the distances in the matrix fits significantly better in a 3D configuration than random arrangements (p<10^{-6}). The co-evolutionary distances between the 45 proteins are thus highly consistent with a 3D representation. Furthermore, the 3D configuration captures 72% of the variation in the co-evolutionary distances (r^2=0.72; raw distances versus embedded distances). In the comparison with the bacterial reference structure, the model of 45 proteins has a slightly lower rmsd (52.6 Å) than the model of only 14 conserved subunits. The extended model clearly recapitulates a number of known elements of the physical structure of the complex (Figure 3). Axis 1 in Figure 3a separates the membrane arm (left) and matrix arm (right) subunits. Among the membrane arm subunits of the Iβ and Iγ sub-complexes (3a, left), axis 2 differentiates between subunits that tend to be encoded by the mitochondrial genome (bottom) and nuclear-encoded (top) subunits. The two subunits NDUFB8 (3a, bottom/right) and NDUFC2 (3b, middle) are located somewhat separate from the remaining Iβ subunits. Among the matrix arm subunits, axis 2 differentiates between the distal N-module subunits V1, V2, and S1 (3a right/top) and the proximal Q-module subunits S7 and S8 (middle), S2, and S3 (bottom). The accessory subunits of hydrophilic Iλ matrix sub-complex (Figure 3, λ) show a tendency towards the top, while those of the Iα sub-complex that are not part of Iλ (3a, α-λ) tend towards the membrane arm subunits at the bottom. The association of this latter group of subunits with the proximal matrix and membrane arm is strongly

Figure 3 Evolutionary model of 38 subunits and seven assembly factors of human complex I. Only the predicted subunit centers are displayed. The evolutionary model was first rotated and scaled to fit to the bacterial structure and both models were then manually rotated into the orientations of the bacterial complex shown in the insets. The symbols code the sub-complex association according to Vogel et al. [93]. Sub-complex Iα-λ represents the fraction of large sub-complex Iα that is not also contained in the smaller matrix arm sub-complex Iλ. Category "?" represents subunits that have not been assigned to any sub-complex. Assembly factors are labeled in italics and red. Subunit names were abbreviated by omitting the "NDUF" or "NADH" prefixes where applicable. N: NUBPL; O7: C20orf7; O38: C8orf38; O56: C2orf56. An interactive three-dimensional model of the evolutionary configuration is provided in Additional file 2.

supported by experimental data (S5 [42,43], A9 [20,44-46], A3 [1,2,47], A6 [48], A8 [48,49], A11 [50], A1 [21]). All assembly factors are located close to the matrix arm subunits. NDUFAF2 (AF2), NDUFAF3 (C3orf60, AF3), C8orf38 (O38), C2orf56 (O56), and NUBPL (N) as well as the permanent subunits A2 and A9 sit on one side of the distal matrix arm core subunits V1, V2, and S1. Only C20orf7 (O7) is placed close to the proximal matrix arm subunits S2, S3 (Figure 3b, right bottom) and the proximal membrane arm subunits A3, A6, and A8. Assembly factor 1 (AF1) is positioned close to the Iλ subunits S8 and A5.

Discussion

Our results show that the evolutionary rates of complex I subunits contain a significant amount of information about the complex's quaternary structure. For the matrix arm we found that about 61% of the correlation in evolutionary rate could be explained by the distances of the subunit centers. This is even more striking if we consider that the evolutionary model was derived from eukaryotic sequences and thus should reflect the matrix arm structure in eukaryotes, while our reference structure is from a bacterium. Indeed, the evolutionary 3D model revealed a twist between the N-module and the Q-module when compared to the bacterial structure, a finding that is supported by experimental data ([6], personal communication).

In the two models that include the membrane arm, mitochondria-encoded subunits were predicted to be separate from nucleus-encoded subunits, which is in line with previous results. The independent variation of

evolutionary rates in the nuclear and mitochondrial genomes [51] may have contributed to the isolation of the mitochondria encoded subunits in our models. Nevertheless, we stress that the position of the membrane arm core subunits, specifically at the proximal end of the matrix arm in both models, indicates a signal of the physical structure in the evolutionary correlation data. Furthermore, the strict separation of the nucleus-encoded Iβ subunits and their mitochondria-encoded counterparts NADH4 and 5 may be explained by other factors, as these two groups also behave differently in experiments [1,52,53]. Interestingly, despite its nuclear encoding, the membrane-integral subunit NDUFA1 of the Iα subcomplex [1,2,54] is positioned close to the membrane arm core, in particular close to NADH2 (Figures 3a and 3b). In *T. thermophilus*, NADH2 is located between the two subunits NADH1 and NADH4 [5] both of which are known physical interactors of NDUFA1 [21]. A direct physical interaction of NDUFA1 with NADH2 is therefore likely.

The evolutionary correlation failed to identify the correct topology of the membrane arm core. A number of biological reasons could explain such a lack of signal. First, long-range structural constraints [5,6] may interfere with the distance-dependent structural constraints that are necessary for a distance-dependent strength of evolutionary rate correlation. Second, the formation of OXPHOS super-complexes with complex I dimers may result in correlations between distant subunits. Indeed, despite their positions at opposite ends of the membrane arm, NADH1 and 5 show a high correlation in evolutionary rates with each other and with subunit CYTb of

complex III [33] consistent with their proximity in OXPHOS complexes organized into respiratory strings [55]. Third, the lack of correlation with physical distance may result from non-adaptive variation in the mitochondria-encoded genes caused by variable and, at least in some eukaryotic taxa, heterogeneous mutation-pressure [56]. Indeed, in a number of animal taxa changes in gene order or mutation-pressure led to non-adaptive changes in mitochondrial genes [57-59]. The mitochondrial genomes of some taxa in our study, such as plants, are clearly different from those in animals (reviewed in [60]) and their genes are likely under different mutation-pressures [61]. Fourth, the embedding of the membrane proteins in two dimensions might reduce the evolutionary constraints to maintain interactions in comparison to proteins that are embedded in three dimensions.

The integration of multiple proteins in a single model assumes that the interactions are permanent and non-competitive. This is clearly not the case for the model of 45 proteins because it includes assembly factors. This model can therefore not exactly represent a physical structure. According to current models, complex I assembles from independent subcomplexes [62]. Of the assembly factors required for this process and included in our study, only NDUFAF1 (AF1) is required for the assembly of the distal membrane arm sub-complex [13,63,64]. In our model, AF1 is located close to the matrix arm, which supports an indirect rather than a direct involvement of AF1 in membrane arm assembly [65]. The distal membrane arm further combines with a pre-formed membrane-anchored proximal matrix/membrane arm that contains the subunits NDUFS2 (S2) and NADH1 (1) and possibly NDUFS3 (S3) and NDUFS7 (S7) [62,64] and whose assembly involves NDUFAF3 (AF3) and possibly C8orf38 (O38) [17,66]. Although the membrane-association of AF3 and O38 is not reflected in our data, they form a tightly co-evolving triple with C2orf56 (O56), which is known to bind the proximal matrix arm subunit S2 [12]. The high correlation in evolutionary rates between AF3, O38, and O56 suggest strong selective constraints on their cooperation during the assembly of the proximal matrix/membrane arm sub-complex. The fourth assembly factor that has been experimentally linked to the proximal membrane arm, C20orf7 (O7) [18,64], is indeed placed close to the proximal matrix arm subunits S2, S3 (Figure 3b, right bottom), and the proximal membrane arm subunits A3, A6, and A8 [2,47,48].

After the joining of the two membrane arm intermediates, the proximal matrix arm is further extended. This step involves the NUBPL-mediated assembly of at least one FeS-cluster into the distal matrix arm [11,67]. In the evolutionary configuration the assembly factor NUBPL

is positioned side by side with the permanent subunit NDUFA2 (A2; Figure 3b, right top). Like NUBPL, A2 is associated with the distal matrix arm [68]. The highly conserved A2 subunit is structurally similar to thioredoxin-like proteins with a loop-region of probably variable conformation that contains two cysteines in human (C24 and C58) [69]. These cysteines can form a revertible disulfide bridge with an *in-vitro* redox-potential in the range of the large majority of isopotential FeS-clusters of complex I [69,70]. Although the cysteines are not fully conserved, occasionally FeS-clusters are bound by serine, histidine, or aspartate [71]. Indeed, the human serine 30 in NDUFA2 is a good candidate for FeS-cluster binding because it is perfectly conserved in all species, with the notable exceptions of *Trypanosoma* and *Leishmania*, in which it is substituted by cysteine. Together these observations and the very strong evolutionary rate correlation of A2 and NUBPL support an involvement of A2 in complex I associated FeS-cluster assembly or maintenance. The peripheral position of A2 and NUBPL in the model could be a consequence of other strong evolutionary constraints not directly related to complex I.

Also NDUFAF2 (AF2, B17.2L) has been linked to the assembly of the distal matrix arm [14,64]. Interestingly, the evolutionary data position AF2 directly besides its paralog NDUFA12 (A12, B17.2) [10,14]. Like AF2, A12 is known to be associated to the distal-matrix arm to which it is directly recruited from the mitochondrial matrix [68]. The correlation in evolutionary rates and the independent co-loss in multiple complex I lacking taxa [10] support an evolutionarily conserved functional relationship of AF2 and A12. It is tempting to speculate that AF2 temporarily binds at the binding site of A12, e.g. to stabilize the local structural context, and is later substituted by its paralog. Such close positioning and physical interaction of homologous proteins within the same protein complex is one of the prevailing trends in the "fate" of duplicated proteins in complexes [72]. Complex I appears to add another twist to this pattern in the sense that the predicted interaction is only temporary.

The rate of protein evolution is influenced by diverse factors [73], in particular expression and general functional relatedness [39,40,74]. It is therefore even more remarkable that we found physical distance to be the major determinant of the evolutionary rate correlation for the complex I matrix arm. However, this result does not apply to the whole complex. Thus, to establish whether the mirror-tree/MDS combination is a good general method to predict quaternary structures, other complexes need to be analyzed. Furthermore, instead of using the mirror-tree method one could use residue correlation to measure the co-evolution of subunits more directly. Residue correlation has been used to predict

contact interfaces for protein pairs [30,75] and to investigate a rotation-symmetric homo-multimeric complex [76]. A simple implementation would be to integrate pairwise residue correlations [28] or correlations that account for indirect correlations [30,31,77] or phylogenetic dependency [76,78] by in-silico two-hybrid [80] into subunit distances and map these into three dimensions by multidimensional scaling.

Conclusions

The correlations of evolutionary rates between subunits of the eukaryotic complex I contain detailed information about the structural arrangement of the matrix arm subunits. This allowed us to make specific predictions about the positions of supernumerary subunits and assembly factors of the matrix arm, which may guide further experimental investigations. Multidimensional scaling could not reconstruct the structure of the membrane arm core. A future analysis will have to investigate what may cause this lack a spatial signal along the membrane arm and thus clarify in particular the relevance of conformational dynamics and super-complex arrangement into a respiratory string for the sequence evolution of complex I.

Methods

Alignments

We included 38 permanent subunits and seven assembly factors of human complex I for which a sufficient number of sequences were available. We collected homologous sequences from the nr database [80] using PSI-BLAST (default parameters). Multiple queries from different species were used whenever PSI-BLAST failed to find known homologs [see Additional file 1]. For A6, B9, A12, and AF2, orthologous groups were manually identified in neighbor-joining trees constructed with identity matrices and correcting for multiple substitutions. Species overlap between the partitions was used to divide the trees into separate orthologous groups. The remaining subunits were treated by a different protocol. First, to ensure a separation of the paralogs NADH2, 4, and 5, we built a set of trusted orthologs of NADH2, 4, and 5 from those sequences that had the best bidirectional hit with the human query using PSI-BLAST. From these seed sequence sets we computed three HMM profiles and sorted the remaining homologs into the orthologous group to which they showed the best profile-alignment [81]. For all sequence sets we selected as single ortholog per species the sequence with the highest NEEDLE score in a pairwise alignment to the human query [82] (default parameters) and/or manual selection based on multiple alignments (MAFFT [83], CLUSTALW [84,85], HMMER [86], HHSEARCH [81]). The kinetoplastida were excluded from our analysis due to their high level of sequence divergence. To gain high quality alignments, we aligned all sequence sets with CLUSTALW and manually fixed misalignments. The manually curated alignments are provided in Additional file 3. Next, we filtered alignment columns with BMGE [37] (–m BLOSUM30 -g 0.50 -b 4), removed sequences that had more than 33% gaps, and restricted the alignments to those species for which we found at least eight subunits of the complex. Of the 43 alignments, 39 had more than 75 sequences and there was no alignment with less than 44 sequences. Finally, we calculated phylogenetic trees using RAXML [38] (Version 7.2.6, PROTGAMMAMTREV for NADH1/2/3/4/4L/5/6, otherwise PROTGAMMAJTT; 4 rate categories) [see Additional file 4]. A single tree was calculated for the concatenated alignment of NADH3, 4L, and 6.

Evolutionary correlation

We calculated evolutionary correlation using a variant of the mirror-tree method [27]. Every subunit's tree was transformed into a vector v containing the pairwise distances between pairs of species in the tree. Because all subunit's trees represent the evolution of proteins within the same species phylogeny they all are similar to that phylogeny and to each other. To remove this basic similarity of the distances we applied the orthogonal projection method developed by Sato *et al.* [35]. The method projects each sequence distance vector v on a reference distance vector p that represents the underlying species phylogeny. Let v_p be the projection of v onto p, then the corrected vector v^* is the residual vector $v - v_p$. The corrected sequence distance vector is thus calculated by

$$v^* = v - v_p = v - \frac{p^T v}{p^T p} p$$

with the row vector p^T [87]. We derived the reference vector p directly from the subunits' distance vectors, as suggested by Kann *et al.* [88]. Specifically, the reference distance between a pair of species was calculated as the average of the distances between these species in the trees of the complex I subunits. Note that this choice of reference as an average of the analyzed vectors also removes the specific pattern of co-variation in evolutionary rates that reflects selective pressure on the complex as a whole. It thus focuses the results on distances between the subunits rather than their distances to unrelated proteins. Finally, the corrected distance vectors were correlated with each other by Spearman rank correlation to yield the subunits' co-evolutionary similarity [27]. We required that the species pairs were present in at least five of our 43 distance vectors. Species pairs occurring in fewer than five vectors were ignored in the correlations. Note that our choice of the set of subunits included the requirement that there are at least 15 species in all pairs of alignments. Only for 17 out of 903 subunit pairs, the correlation values were based on

less than 30 species. The mirror-tree method has the advantages of being easily implemented and it requires low computational resources, even with correction for the basic correlation due to the shared phylogeny.

Multidimensional scaling (cMDS)

The co-evolutionary similarities r were linearly transformed into dissimilarities d by first taking the inverse with respect to the maximum correlation coefficient, i.e. $d'=1-r$, and then rescaling to the interval [0,1] using $d = d' / \max(d')$ [see Additional file 5]. This transformation considers negative correlations in evolutionary rates as negative evidence of physical interaction. We used classical multidimensional scaling (cMDS) as implemented by the R function cmdscale [89] (default parameters) to find the matrix X of coordinates of n points (rows, subunits) in n dimensions (columns) such that the distances between the embedded points are as similar as possible to the original co-evolutionary dissimilarities. Our description of cMDS closely follows that by Borg and Groenen [90]. In detail, for a dissimilarity matrix Δ, cMDS minimizes the loss function $L(X) = ||XX^{T} - B_{\Delta}||^{2}$, where XX^{T} is the scalar product matrix of the embedded coordinates and $B_{\Delta} = -1/2\ J\Delta^{(2)}J$ is the double centered squared dissimilarity matrix with the centering matrix $J = I - n^{-1}\ 11^{T}$, I is the identity matrix, and 1 is a n x 1 matrix of 1s. The solution is found analytically by eigen-decomposition of $B_{\Delta} = Q\Lambda Q$ and calculation of $X = Q_{+}\Lambda_{+}^{1/2}$, where Λ_{+} represents the matrix of the largest k eigenvalues greater than zero and Q_{+} the corresponding columns of Q. The relative magnitudes of the eigenvalues in Λ correspond to the relative contributions of the columns of X in explaining the raw dissimilarities. The goodness-of-fit of the cMDS configuration of n subunits in the k dimensions is quantified by P_k (formula 5.2 in [41]):

$$P_k = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n-1}\lambda_i}$$

where λ_i is the i-th largest eigenvalue of B_{Δ}. Note that the relation between the co-evolutionary dissimilarity and the distance in the cMDS configuration (Shepard diagram, Additional file 1) indicates that the 3D configuration reflects the raw co-evolutionary distances over its whole range.

Superimposition of configurations

The structure of complex I in the *T. thermophilus* served as our reference [5] (PDB:3M9S). We approximated the mass centers of the subunits as the average of x, y, and z coordinates of their C_β atoms (C_α for glycine) [91]. The evolutionary configuration was fitted by rotation and isometric scaling on the bacterial configuration using generalized Procrustes analysis as implemented in the function

GPA of the R package FactoMineR (Version 1.14) [92]. We quantified the difference between the bacterial configuration T and the evolutionary configuration C of n subunits by their root mean square deviation (rmsd)

$$rmsd(T,C) = \sqrt[2]{\frac{\sum_{i=1}^{n}|t_i - c_i|^2}{n}}$$

where $|t_i - c_i|$ is the distance between the bacterial and predicted center of the i-th subunit.

Additional files

Additional file 1: Microsoft Word 97 Document. Table of human complex I members and query sequence identifiers and Shepard diagrams for the three discussed models [1,17] [94-100].

Additional file 2: Zip-compressed FASTA alignment files. Manually curated and unfiltered alignments. The FASTA header lines contain (1) a short sequence identifier consisting of a number and the abbreviated species name and (2) a long sequence identifier with the number written between the genus and epithet of the species name. Sequence gaps are indicated by '-'. Subunits NADH3, NADH4L, and NADH6 were combined ("nadh34L6").

Additional file 3: Zip-compressed New Hampshire eXtended (NHX) tree files. Gene trees.

Additional file 4: Tabulator delimited text file. Matrix of pairwise co-evolutionary distances.

Additional file 5: VRML97 format. Interactive visualization of the evolutionary configuration. The predicted subunit centers are labeled by the abbreviations used in the article and color-coded according to sub-complex membership [see Additional file 1]. Specifically, most subunits are abbreviated by omitting the "NADH" or "NDUF" prefix, with the exception of NUBPL, C20orf7, C8orf38, and C2orf56 that are abbreviated to N, O7, O38, and O56, respectively. The sub-complexes are Iλ (blue), Iα-λ (white), Iγ (yellow), and Iβ (red). Subunits without sub-complex association are shown in purple. You can display the VRML97 file of the configuration using a VRML viewer like Flux Player (Windows; http://mediamachines.wordpress.com/flux-player-and-flux-studio/) or freewrl (Windows, Linux, Mac; http://freewrl.sourceforge.net/).

References

1. Carroll J, Fearnley IM, Shannon RJ, Hirst J, Walker JE: **Analysis of the subunit composition of complex I from bovine heart mitochondria.** *Mol Cell Proteomics* 2003, **2**(2):117–126.

2. Hirst J, Carroll J, Fearnley IM, Shannon RJ, Walker JE: **The nuclear encoded subunits of complex I from bovine heart mitochondria.** *Biochim Biophys Acta* 2003, **1604**(3):135–150.

3. Garcia-Vallve S: **Contribution of each complex of the mitochondrial respiratory chain in the generation of the proton-motive force.** *Biochem Mol Biol Edu* 2004, **32**(1):17–19.

4. Friedrich T, Abelmann A, Brors B, Guénebaut V, Kintscher L, Leonard K, Rasmussen T, Scheide D, Schlitt A, Schulte U, et al: **Redox components and structure of the respiratory NADH:ubiquinone oxidoreductase (complex I).** *Biochim Biophys Acta* 1998, **1365**(1–2):215–219.

5. Efremov RG, Baradaran R, Sazanov LA: **The architecture of respiratory complex I.** *Nature* 2010, **465**(7297):441–445.

6. Hunte C, Zickermann V, Brandt U: **Functional modules and structural basis of conformational coupling in mitochondrial complex I.** *Science* 2010, **329** (5990):448–451.

7. Efremov RG, Sazanov LA: **Structure of the membrane domain of respiratory complex I.** *Nature* 2011, **476**(7361):414–420.

8. Sazanov LA, Hinchliffe P: **Structure of the hydrophilic domain of respiratory complex I from Thermus thermophilus.** *Science* 2006, **311**(5766):1430–1436.

9. Yip C-, Harbour ME, Jayawardena K, Fearnley IM, Sazanov LA: **Evolution of respiratory complex I: 'supernumerary' subunits are present in the α-proteobacterial enzyme.** *J Biol Chem* 2010, **286**(7):5023–5033.

10. Gabaldón T, Rainey D, Huynen MA: **Tracing the evolution of a large protein complex in the eukaryotes, NADH:ubiquinone oxidoreductase (Complex I).** *J Mol Biol* 2005, **348**(4):857–870.

11. Bych K, Kerscher S, Netz DJA, Pierik AJ, Zwicker K, Huynen MA, Lill R, Brandt U, Balk J: **The iron-sulphur protein Ind1 is required for effective complex I assembly.** *EMBO J* 2008, **27**(12):1736–1746.

12. Carilla-Latorre S, Gallardo ME, Annesley SJ, Calvo-Garrido J, Grana O, Accari SL, Smith PK, Valencia A, Garesse R, Fisher PR, et al: **MidA is a putative methyltransferase that is required for mitochondrial complex I function.** *J Cell Sci* 2010, **123**(10):1674–1683.

13. Küffner R, Rohr A, Schmiede A, Krüll C, Schulte U: **Involvement of two novel chaperones in the assembly of mitochondrial NADH:Ubiquinone oxidoreductase (complex I).** *J Mol Biol* 1998, **283**(2):409–417.

14. Ogilvie I, Kennaway NG, Shoubridge EA: **A molecular chaperone for mitochondrial complex I assembly is mutated in a progressive encephalopathy.** *J Clin Invest* 2005, **115**(10):2784–2792.

15. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong S-E, Walford GA, Sugiana C, Boneh A, Chen WK, et al: **A mitochondrial protein compendium elucidates complex I disease biology.** *Cell* 2008, **134**(1):112–123.

16. Saada A, Edvardson S, Rapoport M, Shaag A, Amry K, Miller C, Lorberboum-Galski H, Elpeleg O: **C6ORF66 is an assembly factor of mitochondrial complex I.** *Am J Hum Genet* 2008, **82**(1):32–38.

17. Saada A, Vogel RO, Hoefs SJ, van den Brand MA, Wessels HJ, Willems PH, Venselaar H, Shaag A, Barghuti F, Reish O, et al: **Mutations in NDUFAF3 (C3ORF60), Encoding an NDUFAF4 (C6ORF66)-Interacting Complex I Assembly Protein, Cause Fatal Neonatal Mitochondrial Disease.** *Am J Hum Genet* 2009, **84**:718–727.

18. Sugiana C, Pagliarini DJ, McKenzie M, Kirby DM, Salemi R, Abu-Amero KK, Dahl H-HM, Hutchison WM, Vascotto KA, Smith SM, et al: **Mutation of C20orf7 disrupts complex I assembly and causes lethal neonatal mitochondrial disease.** *Am J Hum Genet* 2008, **83**(4):468–478.

19. Vogel RO, Janssen RJRJ, van den Brand MAM, Dieteren CEJ, Verkaart S, Koopman WJH, Willems PHGM, Pluk W, van den Heuvel LPWJ, Smeitink JAM, et al: **Cytosolic signaling protein Ecsit also localizes to mitochondria where it interacts with chaperone NDUFAF1 and functions in complex I assembly.** *Genes Dev* 2007, **21**(5):615–624.

20. Ugalde C, Vogel R, Huijbens R, Heuvel BVD, Smeitink J, Nijtmans L: **Human mitochondrial complex I assembles through the combination of evolutionary conserved modules: a framework to interpret complex I deficiencies.** *Hum Mol Genet* 2004, **13**(20):2461–2472.

21. Gershoni M, Fuchs A, Shani N, Fridman Y, Corral-Debrinski M, Aharoni A, Frishman D, Mishmar D: **Co-evolution predicts direct interactions between mtDNA and nuclear DNA-encoded subunits of oxidative phosphorylation complex I.** *J Mol Biol* 2010, **404**(1):158–171.

22. Yamaguchi M, Hatefi Y: **Mitochondrial NADH:ubiquinone oxidoreductase (complex I): proximity of the subunits of the flavoprotein and the iron-sulfur protein subcomplexes.** *Biochemistry* 1993, **32**(8):1935–1939.

23. Huynen MA, Bork P: **Measuring genome evolution.** *Proc Natl Acad Sci U S A* 1998, **95**(11):5849–5856.

24. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci U S A* 1999, **96**(8):4285–4288.

25. Gabaldón T, Huynen MA: **Lineage-specific gene loss following mitochondrial endosymbiosis and its potential for function prediction in eukaryotes.** *Bioinformatics* 2005, **21**(Suppl 2):ii144–ii150.

26. Kensche PR, van Noort V, Dutilh B, Huynen MA: **Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution.** *J R Soc Interface* 2008, **5**(19):151–170.

27. Pazos F, Valencia A: **Similarity of phylogenetic trees as indicator of protein-protein interaction.** *Protein Eng* 2001, **14**(9):609–614.

28. Göbel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins.** *Proteins* 1994, **18**(4):309–317.

29. Neher E: **How frequent are correlated changes in families of protein sequences?.** *Proc Natl Acad Sci U S A* 1994, **91**(1):98–102.

30. Burger L, van Nimwegen E: **Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method.** *Mol Syst Biol* 2008, **4**:165.

31. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T: **Identification of direct residue contacts in protein-protein interaction by message passing.** *Proc Natl Acad Sci U S A* 2009, **106**(1):67–72.

32. Huynen MA, de Hollander M, Szklarczyk R: **Mitochondrial proteome evolution and genetic disease.** *Biochim Biophys Acta* 2009, **1792**(12):1122–1129.

33. Yang M, Ge Y, Wu J, Xiao J, Yu J: **Coevolution study of mitochondria respiratory chain proteins: Toward the understanding of protein-protein interaction.** *J Genet Genomics* 2011, **38**(5):201–207.

34. Pazos F, Ranea JA, Juan D, Sternberg MJ: **Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome.** *J Mol Biol* 2005, **352**(4):1002–1015.

35. Sato T, Yamanishi Y, Kanehisa M, Toh H: **The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships.** *Bioinformatics* 2005, **21**(17):3482–3489.

36. Juan D, Pazos F, Valencia A: **High-confidence prediction of global interactomes based on genome-wide coevolutionary networks.** *Proc Natl Acad Sci* 2008, **105**(3):934–939.

37. Criscuolo A, Gribaldo S: **BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments.** *BMC Evol Biol* 2010, **10**(1):210.

38. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688–2690.

39. Hakes L, Lovell SC, Oliver SG, Robertson DL: **Specificity in protein interactions and its relationship with sequence diversity and coevolution.** *Proc Natl Acad Sci U S A* 2007, **104**(19):7999–8004.

40. Clark NL, Alani E, Aquadro CF: **Evolutionary rate covariation reveals shared functionality and coexpression of genes.** *Genome Res* 2012, **22**(4):714–720.

41. Everitt BS, Dunn G: *Applied multivariate data analysis / Brian S. Everitt and Graham Dunn.* London; Melbourne: Edward Arnold; 1991.

42. Marques I, Dencher NA, Videira A, Krause F: **Supramolecular Organization of the Respiratory Chain in Neurospora crassa Mitochondria.** *Eukaryot Cell* 2007, **6**(12):2391–2405.

43. Marques I, Ushakova AV, Duarte M, Videira A: **Role of the Conserved Cysteine Residues of the 11.5 kDa Subunit in Complex I Catalytic Properties.** *J Biochem* 2007, **141**(4):489–493.

44. Fearnley IM, Finel M, Skehel JM, Walker JE: **NADH:ubiquinone oxidoreductase from bovine heart mitochondria. cDNA sequences of the import precursors of the nuclear-encoded 39 kDa and 42 kDa subunits.** *Biochem J* 1991, **278**:821–829.

45. Abdrakhmanova A, Zwicker K, Kerscher S, Zickermann V, Brandt U: **Tight binding of NADPH to the 39-kDa subunit of complex I is not required for catalytic activity but stabilizes the multiprotein complex.** *Biochimica et Biophysica Acta (BBA). Bioenergetics* 2006, **1757**(12):1676–1682.

46. Antonicka H, Ogilvie I, Taivassalo T, Anitori RP, Haller RG, Vissing J, Kennaway NG, Shoubridge EA: **Identification and characterization of a common set of complex I assembly intermediates in mitochondria from patients with complex I deficiency.** *J Biol Chem* 2003, **278**(44):43081–43088.

47. Heinrich H, Azevedo JE, Werner S: **Characterization of the 9.5-kDa ubiquinone-binding protein of NADH:ubiquinone oxidoreductase (complex I) from Neurospora crassa.** *Biochemistry* 1992, **31**(46):11420–11424.

48. Angerer H, Zwicker K, Wumaier Z, Sokolova L, Heide H, Steger M, Kaiser S, Nübel E, Brutschy B, Radermacher M, *et al*: **A scaffold of accessory subunits links the peripheral arm and the distal proton pumping module of mitochondrial complex I.** *Biochem J* 2011, **437**:279–288.

49. da Silva MV, Alves PC, Duarte M, Mota N, da Cunha AL, Harkness TA, Nargang FE, Videira A: **Disruption of the nuclear gene encoding the 20.8-kDa subunit of NADH: ubiquinone reductase of Neurospora mitochondria.** *Mol Gen Genet* 1996, **252**(1–2):177–183.

50. Nehls U, Friedrich T, Schmiede A, Ohnishi T, Weiss H: **Characterization of assembly intermediates of NADH:ubiquinone oxidoreductase (complex I) accumulated in Neurospora mitochondria by gene disruption.** *J Mol Biol* 1992, **227**(4):1032–1042.

51. Lynch M, Koskella B, Schaack S: **Mutation Pressure and the Evolution of Organelle Genomic Architecture.** *Science* 2006, **311**(5768):1727–1730.

52. Sazanov LA, Peak-Chew SY, Fearnley IM, Walker JE: **Resolution of the membrane domain of bovine complex I into subcomplexes: implications for the structural organization of the enzyme.** *Biochemistry* 2000, **39** (24):7229–7235.

53. Lemma-Gray P, Valusová E, Carroll CA, Weintraub ST, Musatov A, Robinson NC: **Subunit analysis of bovine heart complex I by reversed-phase high-performance liquid chromatography, electrospray ionization-tandem mass spectrometry, and matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry.** *Anal Biochem* 2008, **382**(2):116–121.

54. Marques I, Duarte M, Videira A: **The 9.8 kDa Subunit of Complex I, Related to Bacterial Na+–translocating NADH Dehydrogenases, is Required for Enzyme Assembly and Function in Neurospora crassa.** *J Mol Biol* 2003, **329**(2):283–290.

55. Dudkina NV, Kouřil R, Peters K, Braun H-P, Boekema EJ: **Structure and function of mitochondrial supercomplexes.** *Biochim Biophys Acta* 2010, **1797**(6–7):664–670.

56. Reyes A, Gissi C, Pesole G, Saccone C: **Asymmetrical directional mutation pressure in the mitochondrial genome of mammals.** *Mol Biol Evol* 1998, **15**(8):957–966.

57. Delorme MO, Hénaut A: **Codon usage is imposed by the gene location in the transcription unit.** *Curr Genet* 1991, **20**(5):353–358.

58. Asakawa S, Kumazawa Y, Araki T, Himeno H, Miura K, Watanabe K: **Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes.** *J Mol Evol* 1991, **32**(6):511–520.

59. Wei S-J, Shi M, Chen X-X, Sharkey MJ, van Achterberg C, Ye G-Y, He J-H: **New views on strand asymmetry in insect mitochondrial genomes.** *PLoS One* 2010, **5**(9):e12708.

60. Nosek J, Tomaska L: **Mitochondrial genome diversity: evolution of the molecular architecture and replication strategy.** *Curr Genet* 2003, **44**(2):73–84.

61. Nikolaou C, Almirantis Y: **Deviations from Chargaff's second parity rule in organellar DNA: Insights into the evolution of organellar genomes.** *Gene* 2006, **381**:34–41.

62. Lazarou M, Thorburn DR, Ryan MT, McKenzie M: **Assembly of mitochondrial complex I and defects in disease.** *Biochim Biophys Acta, Mol Cell Res* 2009, **1793**(1):78–88.

63. Dunning CJR, McKenzie M, Sugiana C, Lazarou M, Silke J, Connelly A, Fletcher JM, Kirby DM, Thorburn DR, Ryan MT: **Human CIA30 is involved in the early assembly of mitochondrial complex I and mutations in its gene cause disease.** *EMBO J* 2007, **26**(13):3227–3237.

64. Lazarou M, McKenzie M, Ohtake A, Thorburn DR, Ryan MT: **Analysis of the assembly profiles for mitochondrial- and nuclear-DNA-encoded subunits into complex I.** *Mol Cell Biol* 2007, **27**(12):4228–4237.

65. Vogel RO, van den Brand MAM, Rodenburg RJ, van den Heuvel LPWJ, Tsuneoka M, Smeitink JAM, Nijtmans LGJ: **Investigation of the complex I assembly chaperones B17.2L and NDUFAF1 in a cohort of CI deficient patients.** *Mol Genet Metab* 2007, **91**(2):176–182.

66. McKenzie M, Tucker EJ, Compton AG, Lazarou M, George C, Thorburn DR, Ryan MT: **Mutations in the Gene Encoding C8orf38 Block Complex I Assembly by Inhibiting Production of the Mitochondria-Encoded Subunit ND1.** *J Mol Biol* 2011, **414**(3):413–426.

67. Sheftel AD, Stehling O, Pierik AJ, Netz DJA, Kerscher S, Elsasser H-P, Wittig I, Balk J, Brandt U, Lill R: **Human Ind1, an Iron-Sulfur Cluster Assembly Factor for Respiratory Complex I.** *Mol Cell Biol* 2009, **29**(22):6059–6073.

68. Dieteren CEJ, Willems PHGM, Vogel RO, Swarts HG, Fransen J, Roepman R, Crienen G, Smeitink JAM, Nijtmans LGJ, Koopman WJH: **Subunits of Mitochondrial Complex I Exist as Part of Matrix- and Membrane-associated Subcomplexes in Living Cells.** *J Biol Chem* 2008, **283**(50):34753–34761.

69. Brockmann C, Diehl A, Rehbein K, Strauss H, Schmieder P, Korn B, Kühne R, Oschkinat H: **The oxidized subunit B8 from human complex I adopts a thioredoxin fold.** *Structure* 2004, **12**(9):1645–1654.

70. Brandt U: **Energy converting NADH:quinone oxidoreductase (complex I).** *Annu Rev Biochem* 2006, **75**:69–92.

71. Johnson DC, Dean DR, Smith AD, Johnson MK: **Structure, function, and formation of biological iron-sulfur clusters.** *Annu Rev Biochem* 2005, **74**:247–281.

72. Szklarczyk R, Huynen M, Snel B: **Complex fate of paralogs.** *BMC Evol Biol* 2008, **8**(1):337.

73. Pál C, Papp B, Lercher MJ: **An integrated view of protein evolution.** *Nat Rev Genet* 2006, **7**(5):337–348.

74. Drummond DA, Raval A, Wilke CO: **A single determinant dominates the rate of yeast protein evolution.** *Mol Biol Evol* 2006, **23**(2):327–337.

75. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H: **High-resolution protein complexes from integrating genomic information with molecular simulation.** *Proc Natl Acad Sci U S A* 2009, **106**(52):22124–22129.

76. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M: **Direct-coupling analysis of residue coevolution captures native contacts across many protein families.** *Proc Natl Acad Sci U S A* 2011, **108**(49):E1293–E1301.

77. Sreekumar J, ter Braak CJ, van Ham RC, van Dijk AD: **Correlated mutations via regularized multinomial regression.** *BMC Bioinforma* 2011, **12**:444.

78. Burger L, van Nimwegen E: **Disentangling direct from indirect co-evolution of residues in protein alignments.** *PLoS Comput Biol* 2010, **6**(1):e1000633.

79. Pazos F, Valencia A: **In silico two-hybrid system for the selection of physically interacting protein pairs.** *Proteins* 2002, **47**(2):219–227.

80. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2006, **35**(suppl 1):D5–D12.

81. Söding J: **Protein homology detection by HMM–HMM comparison.** *Bioinformatics* 2005, **21**(7):951–960.

82. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16**(6):276–277.

83. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program.** *Brief Bioinform* 2008, **9**(4):286–298.

84. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**(22):4673–4680.

85. Li K-B: **ClustalW-MPI: ClustalW analysis using distributed and parallel computing.** *Bioinformatics* 2003, **19**(12):1585–1586.

86. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**(9):755–763.

87. Strang G: *Introduction to Linear Algebra*. In.: Wellesley-Cambridge Press; 1998.

88. Kann MG, Shoemaker BA, Panchenko AR, Przytycka TM: **Correlated evolution of interacting proteins: looking behind the mirrortree.** *J Mol Biol* 2009, **385**(1):91–98.

89. R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2011.

90. Borg I, Groenen PJF: *Modern Multidimensional Scaling*. Springer: Theory and Applications; 2005.

91. Singer MS, Vriend G, Bywater RP: **Prediction of protein residue contacts with a PDB-derived likelihood matrix.** *Protein Eng* 2002, **15**(9):721–725.

92. Lê S, Josse J, Husson F: **FactoMineR: An R Package for Multivariate Analysis.** *J Stat Softw* 2008, **25**(1):1–18.

93. Vogel RO, Smeitink JAM, Nijtmans LGJ: **Human mitochondrial complex I assembly: a dynamic and versatile process.** *Biochim Biophys Acta* 2007, **1767**(10):1215–1227.

94. Carroll J, Shannon RJ, Fearnley IM, Walker JE, Hirst J: **Definition of the nuclear encoded protein composition of bovine heart mitochondrial complex I. Identification of two new subunits.** *J Biol Chem* 2002, **277** (52):50311–50317.

95. Fearnley IM, Carroll J, Shannon RJ, Runswick MJ, Walker JE, Hirst J: **GRIM-19, a cell death regulatory gene product, is a subunit of bovine mitochondrial NADH:ubiquinone oxidoreductase (complex I).** *J Biol Chem* 2001, **276** (42):38345–38348.

96. Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, *et al*: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437**(7062):1173–1178.
97. Murray J, Taylor SW, Zhang B, Ghosh SS, Capaldi RA: **Oxidative Damage to Mitochondrial Complex I Due to Peroxynitrite.** *J Biol Chem* 2003, **278** (39):37223–37230.
98. Mishmar D, Ruiz-Pesini E, Mondragon-Palomino M, Procaccio V, Gaut B, Wallace DC: **Adaptive selection of mitochondrial complex I subunits during primate radiation.** *Gene* 2006, **378**:11–18.
99. Nouws J, Nijtmans L, Houten SM, van den Brand M, Huynen M, Venselaar H, Hoefs S, Gloerich J, Kronick J, Hutchin T, *et al*: **Acyl-CoA dehydrogenase 9 is required for the biogenesis of oxidative phosphorylation complex I.** *Cell Metab* 2010, **12**(3):283–294.
100. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P: **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.** *Nucleic Acids Res* 2011, **39**(Database issue):D561–568.