Supporting Information.

Functions of Essential Genes and a Scale-free Protein Interaction Network Revealed by Structure-based Function and Interaction Prediction for a Minimal Genome

Chengxin Zhang[1,‡], Wei Zheng[1,‡], Micah Cheng[2], Gilbert S. Omenn[1,3], Peter L. Freddolino[4,1,], Yang Zhang[1,4,*]*

1 Department of Computational Medicine and Bioinformatics,

2 Department of Electrical Engineering and Computer Science,

3 Departments of Internal Medicine and Human Genetics and School of Public Health,

4 Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109 USA

*(P.L.F.) Email: petefred@umich.edu. (Y.Z.) Email: zhng@umich.edu.

Text S1. Deep learning based contact prediction in C-I-TASSER.

Figure S1. Violin plots for portions of residues predicted by TMHMM2.0 to be within transmembrane helices (y-axis) for JCVI-syn3.0 proteins that are annotated (left) versus unannotated (right) by C-I-TASSER/COFACTOR with C-score>0.5 for specific GO terms in the MF **(A)**, BP **(B)** and CC **(C)** aspects.

Figure S2. Structural alignment (see text for details) of the predicted structure of MMSYN1_0877 with the nine closest structural homologs identified in the Protein Data Bank (one candidate, PDB ID 4M58, is excluded from the alignment due to the presence of a large insertion). The structures shown here represent substrate binding domains for ECF systems targeting riboflavin (5KBW, 3P5N), biotin (4DVE), thiamine (4TKR, 3RLB, 5EDL), folate (4HUQ), pantothenate (4RFS), and hydroxymethyl pyridine (4HZU). Positions which are within 0.5 nm of the riboflavin in the 3P5N structure are highlighted with arrows. Note that only segments of the protein surrounding those contact points are shown. Numbering corresponds to the PDB structures rather than the original protein sequences; the ruler along the top shows positions in the 3P5N sequence.

Figure S3. A random PPI network for syn3.0, where 2483 of all 95703 protein pairs are randomly selected as the positive PPI pairs.

Text S1. Deep learning based contact prediction in C-I-TASSER. The C-I-TASSER pipeline incorporates residue-residue contact predicted from a multiple sequence alignment (MS)A by three in-house deep learning algorithms ResPRE [1], ResTriplet [2], and TripletRes [2]. All three predictors sequence features derived from direct coupling analysis (DCA), which quantifies the co-evolution among amino acid positions in the MSA, and feed these features to deep residual convolutional neural networks (ResNets); but they differ in technical details of how the DCA features are generated and combined by ResNets.

In ResPRE, the DCA features are computed as a "precision matrix" (PRE), which is similar to an inverse covariance matrix[3] but regularized by the L2 norm rather than the L1 norm. This precision matrix is input to ResNets to predict an $L*L$ contact map, where L is the length of target protein. In this contact maps, 1 means a pair of residue is in contact, i.e. distance between Cβ atoms <8Å, while a 0 in the contact map means the residue pair not in contact.

ResTriplet includes not only ResPRE but also another two in-house predictors, ResPLM and ResCOV. These two predictors have the same ResNet architecture as ResPRE, but the DCA features are calculated by pseudo-likelihood estimation (PLM) [4] and covariance (COV) [5], respectively. The output of these three ResNet predictors are used as input features for a fourth ResNet, which outputs the final contact map. The four ResNets are trained one-by-one.

TripletRes uses the same set of PRE, PLM, and COV features as ResTriplet. Each of the three sets of features also correspond to a one ResNet, whose outputs are combined by the fourth ResNet to derive the final contact map. However, instead of training the four ResNets separately as in ResTriplet, TripletRes train all four predictors together in an end-to-end

fashion. As shown in the recent 13th Critical Assessment of protein Structure Prediction (CASP13) community-wide experiment [2], ResTriplet slightly outperforms TripletRes for easy targets with more homologs while TripletRes outperforms ResTriplet for hard targets with less homologs; both ResTriplet and TripletRes outperform ResPRE.

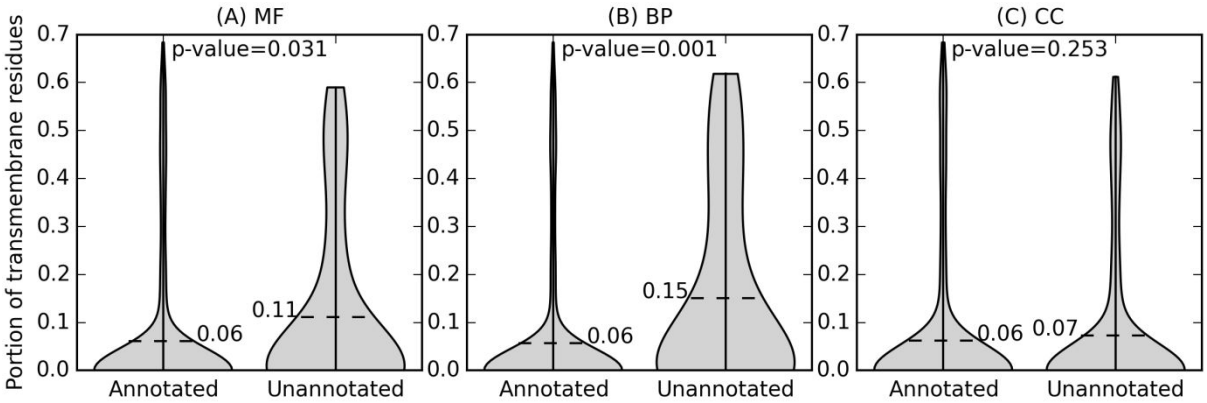

Figure S1. Violin plots for portions of residues predicted by TMHMM2.0 to be within transmembrane helices (*y*-axis) for JCVI-syn3.0 proteins that are annotated (left) versus unannotated (right) by C-I-TASSER/COFACTOR with C-score>0.5 for specific GO terms in the MF **(A)**, BP **(B)** and CC **(C)** aspects. The *p*-value is calculated by single-tailed unpaired t-test to test if the average portion of transmembrane residues (dashed lines) for C-I-TASSER/COFACTOR annotated proteins is significantly smaller than that for unannotated proteins.

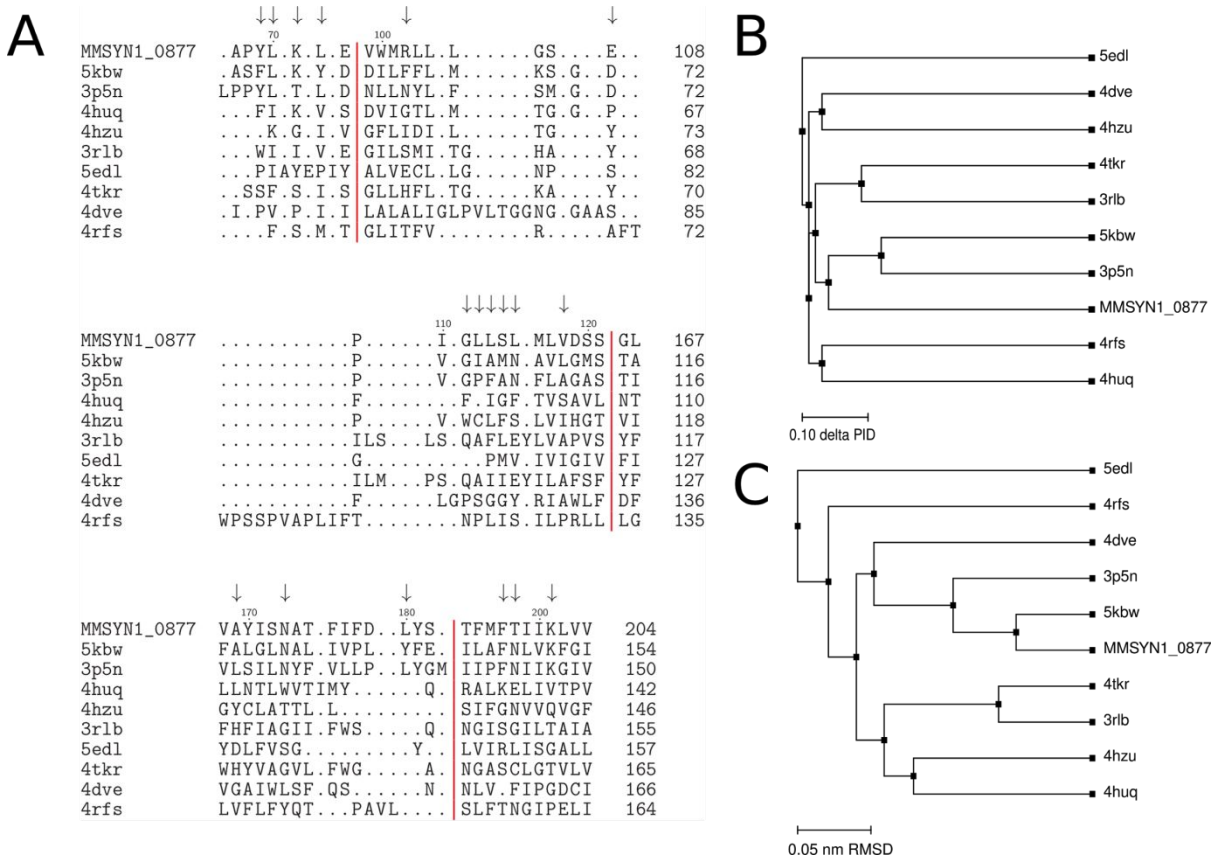

Figure S2. MMSYN1_0877 structurally resembles known riboflavin ECF substrate binding domains. **(A)** Structural alignment (see text for details) of the predicted structure of MMSYN1_0877 with the nine closest structural homologs identified in the Protein Data Bank (one candidate, PDB ID 4M58, is excluded from the alignment due to the presence of a large insertion). The structures shown here represent substrate binding domains for ECF systems targeting riboflavin (5KBW, 3P5N), biotin (4DVE), thiamine (4TKR, 3RLB, 5EDL), folate (4HUQ), pantothenate (4RFS), and hydroxymethyl pyridine (4HZU). Positions which are within 0.5 nm of the riboflavin in the 3P5N structure are highlighted with arrows. Note that only segments of the protein surrounding those contact points are shown. Numbering corresponds to the PDB structures rather than the original protein sequences; the ruler along the top shows positions in the 3P5N sequence. **(B)** UPGMA clustering of the proteins shown in panel A using percent identity at structurally aligned columns as a distance measure. **(C)** UPGMA clustering of the proteins shown in panel A using structural RMSD as a distance measure.

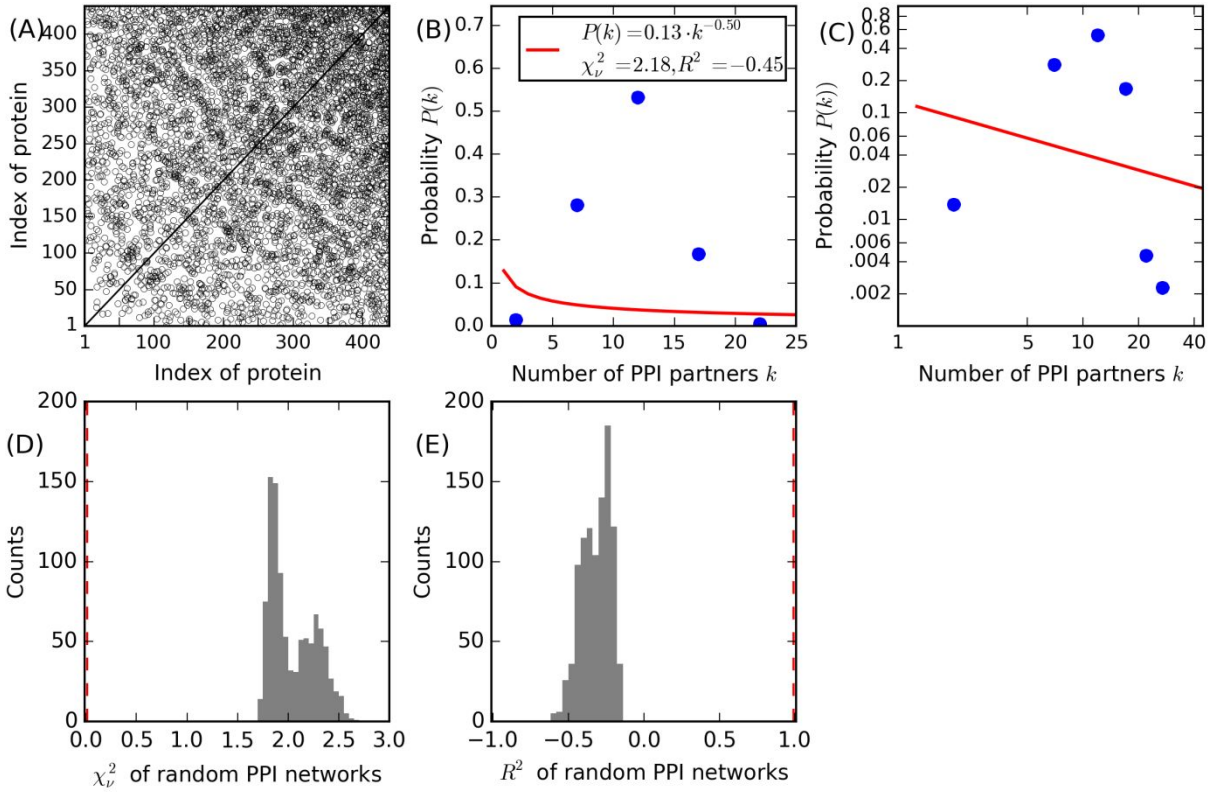

Figure S3. A random PPI network for syn3.0, where 2483 of all 95703 protein pairs are randomly selected as the positive PPI pairs. The number of positive pairs in this random network is therefore identical to the SPRING-predicted PPI network shown in **Figure 1**. **(A)** Scatter plot of PPIs for all syn3.0 proteins ranked in ascending number of PPI partners, where a point means the protein pair is predicted to have a PPI. **(B-C)** Observed distribution (circles) for the number of PPI partners per protein in linear **(B)** and log **(C)** scale, and the power law fit (lines). In the inset, χ^2_ν is the reduced chi-squared statistic (lower values are better, with 0 being a perfect fit) and R^2 is the coefficient of determination (the higher the better, with 1 being a perfect fit), respectively, to quantify the goodness of fit. Both metrics indicate that power law fits poorly to the distribution of the number of PPI partners per protein. **(D-E)** Histogram of χ^2_ν **(D)** and R^2 **(E)** values of 1000 randomly generated PPI networks for syn3.0 with the same number of positive pairs as the SPRING-predicted network. The vertical dash lines to the left **(D)** or right **(E)** of the histograms indicates χ^2_ν =0.01 and R^2=0.99, respectively, in the SPRING-predicted network (**Figure 4B**), which is consistently better fitted to a power law distribution to all 1000 randomly generated PPI networks.

REFERENCE

(1) Li, Y.; Hu, J.; Zhang, C.; Yu, D.-J.; Zhang, Y. ResPRE: High-Accuracy Protein Contact Prediction by Coupling Precision Matrix with Deep Residual Neural Networks. *Bioinformatics* **2019**, *35* (22), 4647–4655.

(2) Li, Y.; Zhang, C.; Bell, E. W.; Yu, D.; Zhang, Y. Ensembling Multiple Raw Coevolutionary Features with Deep Residual Neural Networks for Contact-map Prediction in CASP13. *Proteins* **2019**, *87* (12), 1082–1091.

(3) Jones, D. T.; Buchan, D. W. A.; Cozzetto, D.; Pontil, M. PSICOV: Precise Structural Contact Prediction Using Sparse Inverse Covariance Estimation on Large Multiple Sequence Alignments. *Bioinformatics* **2012**, *28* (2), 184–190.

(4) Seemayer, S.; Gruber, M.; Söding, J. CCMpred—fast and Precise Prediction of Protein Residue–residue Contacts from Correlated Mutations. *Bioinformatics* **2014**, *30* (21), 3128–3130.

(5) Jones, D. T.; Kandathil, S. M. High Precision in Protein Contact Prediction Using Fully Convolutional Neural Networks and Minimal Sequence Features. *Bioinformatics* **2018**, *34* (19), 3308–3315.

(6) Breuer, M.; Earnest, T. M.; Merryman, C.; Wise, K. S.; Sun, L.; Lynott, M. R.; Hutchison, C. A.; Smith, H. O.; Lapek, J. D.; Gonzalez, D. J.; de Crécy-Lagard, V.; Haas, D.; Hanson, A. D.; Labhsetwar, P.; Glass, J. I.; Luthey-Schulten, Z. Essential Metabolism for a Minimal Cell. *Elife* **2019**, *8*. https://doi.org/10.7554/eLife.36842.