SUPPLEMENTARY MATERIAL

Elements of Coevolution in Biological Sequences

Olivier Rivoire
CNRS, LIPhy UMR 5588, Grenoble, France

Preprocessing of the alignment

As input for the identification of sectors and sectons in the trypsin family, we downloaded the full alignment PF00089 from Pfam (version 26.0, Nov. 2011) [1]. This alignment contains $M_0 = 14720$ sequences. It is represented by an array x^a_{si} where s labels the sequences (row in the alignment), i the positions (columns) and a is a number between 1 and 20 (each number is associated with one of the $A = 20$ amino acids); $x^a_{si} = 1$ indicates that sequence i has amino acid a at position i, and $x^a_{si} = 0$ otherwise. As a reference for truncating the alignment and comparing to structural data, we used the sequence and structure of rat trypsin, chain E of the PDB id 3TGI in the Protein Data Base [2], which consists of $L_0 = 223$ positions.

To clean the alignment from an excess of gaps, the following operations were performed:
(1) Truncation of positions based on the reference sequence. As the alignment does not contain the last 7 positions of the reference sequence, this step leaves $L_1 = 216$ positions.
(2) Removal of sequences with a fraction of gaps exceeding $\gamma_{\rm seq} = 0.2$, or with a sequence similarity to the reference sequence below $s_{\rm min} = 0.2$, where sequence similarity is defined by

$$S^{(1)}_{rs} = \frac{1}{L_1}\sum_{i=1}^{L_1}\sum_{a=1}^{20} x^a_{ri}x^a_{si}. \tag{1}$$

This step leaves $M = 9589$ sequences.
(3) Removal of positions with a fraction of gaps exceeding $\gamma_{\rm pos} = 0.2$. The frequencies of gaps are computed with sequence weights as defined by Eq. (2), using $S^{(1)}_{rs}$. This step leaves $L = 204$ positions.

The parameters $\gamma_{\rm seq} = 0.2$, $\gamma_{\rm pos} = 0.2$ and $s_{\rm min} = 0.2$ are chosen to mitigate the effects of gaps, but the results are not sensitive to their exact values. A more in-depth analysis of the structure of sequence correlations can reveal further information, and may suggest the removal of additional sequences, but this analysis is beyond the scope of the present study [3].

Sequence-weighted frequencies

Following Ref. [4], the uneven sampling of sequences is alleviated by introducing sequence weights defined by

$$w_s \equiv \frac{\nu_s^{-1}}{\sum_r \nu_r^{-1}}, \quad \text{with} \quad \nu_s \equiv |\{r : S_{rs} > \delta\}|, \tag{2}$$

i.e., ν_r counts the number of sequences r within distance δ of sequence s, where the distance between two sequences is the fraction of amino acids by which they differ,

$$S_{rs} = \frac{1}{L}\sum_{i=1}^{L}\sum_{a=0}^{20} \tilde{x}^a_{ri}\tilde{x}^a_{si}, \tag{3}$$

where here, for numerical convenience, we include gaps by defining $\tilde{x}^a_{si} = x^a_{si}$ for $a > 0$ and $x^0_{si} = 1$ if sequence s has a gap at position i. $M' = \sum_r \nu_r^{-1}$ can be interpreted as an effective number of sequences in the alignment. Here we take $\delta = 0.8$, which results in $M' \simeq 4600$ effective sequences.

The sequence weights are used to define the frequency f_i^a of amino acid a at position i and the joint frequency f_{ij}^{ab} of amino acids a, b at positions i, j as

$$f_i^a \equiv \sum_s w_s x_{si}^a, \qquad f_{ij}^{ab} \equiv \sum_s w_s x_{si}^a x_{sj}^b. \tag{4}$$

From these frequencies, a covariance matrix C_{ij}^{ab} is defined by

$$C_{ij}^{ab} = f_{ij}^{ab} - f_i^a f_j^b. \tag{5}$$

Regularization

Direct coupling analysis (DCA) involves the inverse of a covariance matrix but C_{ij}^{ab} is typically non invertible. A regularized covariance matrix is therefore introduced [7], which is defined as

$$\bar{C}_{ij}^{ab} = \bar{f}_{ij}^{ab} - \bar{f}_i^a \bar{f}_j^b, \tag{6}$$

from the regularized frequencies

$$\bar{f}_i^a = (1-\mu) f_i^a + \mu \frac{1}{A+1}, \qquad \bar{f}_{ij}^{ab} = (1-\mu) f_{ij}^{ab} + \mu \frac{1}{(A+1)^2}, \tag{7}$$

where $A = 20$ for multiple sequence alignment (corresponding to the number of amino acids, with $A + 1 = 21$ accounting for gaps), and $A = 1$ for gene co-occurrence data. Here as in Ref. [4], $\mu = 1/2$ is used, which effectively amounts to adding M random sequences to the alignment.

SCA matrix $\mathcal{C}_{ij}$

Statistical Coupling Analysis (SCA) aims at identifying groups of positions under selection for a common functional property, based on two principles: the conservation of amino acids involved in the function, and their correlations induced by cooperative interactions [5]. SCA takes a heuristic approach to combine these two principles by weighting the covariance matrix C_{ij}^{ab} of Eq. (5) with a measure of amino acid conservation W_i^a, defined by

$$W_i^a = \left| \ln \left(\frac{f_i^a (1-q^a)}{(1-f_i^a) q^a} \right) \right|, \tag{8}$$

where $q^a = \sum_{i=1}^L f_i^a / L$ is the mean frequency of amino acid a (close in this case to the mean frequency inferred from a larger set of protein sequences used in Refs. [5, 11]). Taking $W_a^i W_b^j C_{ij}^{ab}$ defines a conservation-weighted correlation matrix, which is reduced to a $L \times L$ correlation matrix between positions as follows [3, 5]:

$$\mathcal{C}_{ij} = \sqrt{\sum_{a,b=1}^{A} \left(W_i^a W_j^a C_{ij}^{ab} \right)^2}. \tag{9}$$

For comparison with DCA, an equivalent quantity $\bar{\mathcal{C}}_{ij}$ is defined with same W_i^a but the regularized covariance matrix $\bar{C}_{ij}^{ab}$ of Eq. (6) instead of C_{ij}^{ab}.

Matrix of direct information $\mathcal{D}_{ij}$

In contrast to SCA, Direct Coupling Analysis (DCA) aims at identifying structural contacts between positions by inferring direct interactions from indirect correlations. It proceeds by a mapping to the problem of reconstructing the couplings J_{ij}^{ab} of a $(A+1)$-state Potts model given its correlations C_{ij}^{ab}. Solving exactly this problem is computationally prohibitive, but mean-field approximations provide a range of alternatives [6]. The simplest of these approximations consists in taking $J = -C^{-1}$. As the matrix C is generally not invertible, the regularized covariance matrix $\bar{C}$ of

Eq. (6) is substituted for C. The couplings J_{ij}^{ab} define for $i \neq j$ a model for the distribution of amino acids at every pair of positions ij,

$$g_{ij}^{ab} = \exp(J_{ij}^{ab} + h_i^a + h_j^b + h_0), \tag{10}$$

where h_i^a, h_j^b, h_0 are uniquely determined by requiring that $\sum_b g_{ij}^{ab} = \bar{f}_i^a$. From g_{ij}^{ab}, a matrix of direct information [4] is defined by

$$\mathcal{D}_{ij} = \sum_{a,b=0}^{A} g_{ij}^{ab} \ln \frac{g_{ij}^{ab}}{\bar{f}_i^a \bar{f}_j^b} \tag{11}$$

where the sum includes $a = 0$ for gaps.

Rotation by independent component analysis (ICA)

Different implementations of ICA use different measures of independence and different algorithms for optimizing them. Here, we use one of the simplest implementations of ICA, called infomax [9] (with modifications introduced in Ref. [10]). We take as input the top k eigenvectors of the correlation matrix $\mathcal{C}_{ij}$ or $\tilde{\mathcal{D}}_{ij}$, which we concatenate in a $k_{\rm top} \times L$ matrix Z (at variance with usual implementations of ICA which take the dataset X as input). The algorithm iteratively updates an unmixing matrix W, starting from the $k_{\rm top} \times k_{\rm top}$ identity matrix $W_0 = I_{k_{\rm top}}$, with increments ΔW given by

$$\Delta W = \eta \left(I_{k_{\rm top}} + \left(1 - \frac{2}{1 + \exp(-WZ)} \right) (WZ)^\top \right) W. \tag{12}$$

The parameter η is a learning rate that has to be sufficiently small for the iterations to converge. In this study, $\eta = 10^{-4}$ led to convergence after 10^4 iterations in applications to the trypsin alignment, and $\eta = 10^{-5}$ after 10^5 iterations in applications to the co-occurrence of genes in bacterial genomes.

The independent components $V^{(k)}$ are obtained by applying W to the eigenvectors in Z. To set their overall scale and sign, we normalize them to unit length ($\sum_i (V_i^{(k)})^2 = 1$) and orient them so that the position i with largest $|V_i^{(k)}|$ satisfies $V_i^{(k)} > 0$. The order of the independent components, which is not necessarily prescribed in other implementations of ICA, is here well defined by the algorithm and is related to the order of the principal components.

Threshold $k_{\rm top}$ in defining sectors

The spectrum of $\mathcal{C}_{ij}$, displayed in Fig. S2, indicates that between 3 to 7 eigenvalues are emerging from a bulk of small eigenvalues. This estimation is confirmed by comparing with the spectra of randomized alignments, where the amino acids are drawn independently at each position i according the frequencies f_i^a, so as to remove the correlations but preserve the distribution of amino acids at each position.

In the main text, we presented the results when selecting $k_{\rm top} = 4$ modes. A smaller number of components may prevent the discrimination between sectors, as shown in Fig. S3, where taking $k_{\rm top} = 3$ causes the red and blue sectors to appear along a same component. Reciprocally, a larger number of components may lead to the splitting of a sector into disconnected subsets, as shown in Fig. S4, where taking $k_{\rm top} = 5$ decomposes the purple sector along two components. In this case, the two components do not define two new sectors, but indicate a partition of the sector into a core and a periphery, as shown in Fig. S5.

Threshold ϵ in defining sectors

Besides the threshold $k_{\rm top}$ for the eigenvalues, the definition of sectors involves a threshold ϵ determining which positions contribute significantly to each component. Here again, ϵ can be estimated from a comparison with randomized alignments, but it is more interesting to notice that several values of ϵ are consistent with structurally

connected sectors. Varying ϵ thus defines a hierarchy of structurally connected positions, from the core of the most conserved positions for large ϵ to a periphery of less conserved positions for smaller ϵ.

As an illustration of this feature, we show in Fig. S6 how the connectivity of each sector, measured by the relative size of its largest structurally connected subset, varies with ϵ. With the possible exception of the cyan sector, the sectors are found to be significantly structurally connected for nearly all values of ϵ. The significance of this finding is assessed by comparing with randomly-formed groups of positions, or with the positions ordered by their overall degree of conservation (Fig. S6(C)).

Composition of sectors

Tab. SI reports the exact composition of the 4 sectors defined in the main text. The green and red sectors have very significant overlap with the green and red sectors previously defined in Ref. [11]. On the other hand, the purple and cyan sectors have only limited overlap with the blue sector defined in this previous study. As a visual representation of the relation between these two definitions, Fig. S7 reproduces Fig. 1, but with colors corresponding to the 3 sectors defined in Ref. [11]. While the implementation of SCA is slightly different in the two studies, the difference mainly stems from the difference between the alignments that serve as inputs: the present study is based on an alignment with nearly 10 more sequences (and $L = 204$ positions instead of 223).

In Fig. 1, we eluded the discussion of positions at the intersection between different $\mathcal{S}_k = \{i : V_i^{(k)} > \epsilon\}$ by excluding them from the definition of sectors. These few positions are however structurally meaningful: Fig. S1 shows that they are at the periphery of one of the two sectors, and, in several instances, at the structural interface between them.

Contacts and sectons

Fig. S8 reports the performance of DCA for predicting contacts for the alignment under study: up to rank 79, all top pairs of positions in terms of $\tilde{\mathcal{D}}_{ij}$ are actual contacts. Tab. SII gives the list of the top 120 pairs and indicates the secton to which they belong, when applicable. Tab. SIII gives the composition of the top 120 sectons and indicates which are connected. Fig. S10 shows the top 36 sectons on the three-dimensional structure of trypsin: the 36th secton is the first to include a position disconnected from the others.

Fig. S11 shows that the top 35 sectons are structurally connected (and slightly more if considering a more stringent cutoff ϵ). It also shows that, as for sectors, the results are insensitive to the choice of the threshold ϵ used in defining the positions contributing significantly to a component. In all figures involving sectons, $k_{\text{top}} = 120$, except Fig. S12, which consider varying k_{top}: the definition of sectons is found to be insensitive to the value of k_{top}, except for the first secton, which appears to be peculiar and unstable. For instance, the first secton is the only secton of size 2 that is not in the top pairs of $\tilde{\mathcal{D}}_{ij}$, and is responsible for the blue curve not reaching 1 in Fig. 2(C).

Not truncating $\mathcal{D}_{ij}$ to $\tilde{\mathcal{D}}_{ij}$ leads to the sectons shown in Fig. S9. Many of these sectons consist of consecutive positions. These trivial sectons are induced by gaps, which tend to be consecutive along the sequence (this feature is partly a consequence of the multiple sequence alignment algorithm, which have a penalty for opening new gaps). Truncating $\mathcal{D}_{ij}$ (or equivalently J_{ij}^{ab}) is a simple if not optimal way of getting rid of these trivial correlations.

Figs. 2(C-D) show that the top pairs of $\tilde{\mathcal{D}}_{ij}$ and sectons contain overlapping but non-equivalent informations. This is also shown by Tab. IV, which analyzes the connected subgraphs of the contact graph inferred from the top pairs of $\tilde{\mathcal{D}}_{ij}$: these connected subgraphs are related, but non-identical to the sectons.

Orthogonal decomposition of the correlations

As in the main text, $\bar{C} = \sum_k |k\rangle\lambda_k\langle k|$ denotes the spectral decomposition of the regularized covariance $\bar{C}$ in the bra-ket notation, with ordered eigenvalues $\lambda_1 \geq \ldots \lambda_L$. This decomposition is used to defined the two matrices

$$\bar{C}^+ = \sum_{k \leq k^*} |k\rangle\lambda_k\langle k|, \quad \text{and} \quad \bar{C}^- = \sum_{k > k^*} |k\rangle\lambda_k\langle k|. \tag{13}$$

Fig. S14 shows the differences between the components of the SCA matrices defined from C (the original covariance matrix with $\mu = 0$ and $k^* = 0$), $\bar{C}$ (the regularized covariance matrix with $\mu = 1/2$ and $k^* = 0$) and $\bar{C}^+$ (with $\mu = 1/2$ and $k^* = 100$). Keeping only the top k^* modes has no incidence on the top components of the SCA matrix used to define sectors, while regularizing has only a minor effect.

The sectons defined from $J^- = -\sum_{k>k^*} |k\rangle\lambda_k^{-1}\langle k|$, with again $k^* = 100$, are comparable to those defined when including all the modes, as shown in Fig. S15 (with the first secton forming an exception, as in Fig. S12). Fig. S16 shows that truncating the top k^* modes of $\bar{C}$ can actually cause more top pairs and more sectons to be structurally connected.

Co-occurrence of orthologous genes in bacterial genomes

Sequenced bacterial genomes and COG annotations were downloaded from NCBI. The initial dataset contained $M_0 = 1432$ genomes and $L_0 = 4467$ COGs.

The following cleaning steps were conducted:
(1) Removal of 'exceptional' genomes with size below 500 kpb or with no less than 60 % of genes annotated by COGs. This step leaves $M = 1108$ genomes.
(2) Removal of COGs that are present in less than $\gamma_c = 0.4$ of the genomes, where gene frequencies are computed with sequence weights using $\delta = 0.9$. The relatively high value $\gamma_c = 0.4$ is meant to reduce the data to a size that is easily tractable computationally; here it leaves $L = 1474$ COGs. Conversely, the choice of $\delta = 0.9$ is meant to preserve a relatively high effective number of genomes, here $M' = 380$. The exact values of these parameters are however not crucial.

The data is represented by a $M \times L$ binary array x_{si} with $x_{si} = 1$ if genome s has at least one gene in COG i, and 0 otherwise. The average occurrence of genes is $q = \sum_{si} x_{si}/(ML) = 0.67$. This dataset is in no way meant to be optimal, and finer definitions of orthology are possible. Our point here is to show that sectors and sectons can be unraveled even from a relatively crude construction of bacterial phylogenetic profiles, leaving for future work the study of more elaborated datasets.

Genomic sectons are obtained exactly as protein sectons, except that the alphabet is now binary ($A = 1$), sequence weights are computed with $\delta = 0.9$, and $\mathcal{D}_{ij}$ is not truncated ($\Delta = 0$, with $\mathcal{D}_{ii} = 0$). The covariance matrix $C_{ij} = f_{ij} - f_i f_j$ is computed from the frequencies f_i and f_{ij} of occurrence of genes and co-occurrence of pairs of genes. The content of first 100 sectons (obtained with $k_{\rm top} = 120$) is reported in Tab. SV, with further details for the top 24 sectons provided in Tab. SVI.

Fig. S20 reports the distribution of sizes of genomic sectons and Fig. S20 analyzes the differences of content between top pairs of $\mathcal{D}_{ij}$, top pairs of C_{ij}, and sectons. Finally, Fig. S18 shows that applying ICA to the top $k_{\rm top} = 120$ eigenvectors of C_{ij} does not lead to localized components as it does when applying it to $\mathcal{D}_{ij}$ (Fig. S17).

Genomic sectors can also be defined following the methods for defining protein sectors. Fig. S21 shows the counterpart of Fig. 1, using here $k_{\rm top} = 6$. In absence of a counterpart for the experimentally determined three-dimensional structure of proteins, assessing the relevance of these sectors is not obvious. Using the partition of COGs into 3 broad functional classes [12], we nevertheless find that 4 of the 6 components support groups of COGs that are significantly enriched in some of these classes, as reported in Tab. SVII. This suggests that genome sectors may be defined as well, which contain co-functional and therefore coevolving genes.

[1] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, et al., Nucl Acids Res **40**, D290 (2011).
[2] A. Pasternak, D. Ringe, and L. Hedstrom, Prot. Science **8**, 253 (1999).
[3] O. Rivoire and R. Ranganathan, (in preparation).
[4] M. Weigt, R. White, H. Szurmant, J. Hoch, and T. Hwa, Proc Nat Acad Sc **106**, 67 (2009).
[5] S. Lockless and R. Ranganathan, Science **286**, 295 (1999).

[6] Y. Roudi, E. Aurell, and J. A. Hertz, Frontiers Comput. Neuro. **3** (2009).
[7] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. **108**, E1293 (2011).
[8] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, PLoS ONE **6**, e28766 (2011).
[9] A. Bell and T. Sejnowski, Neural Comp. **7**, 1129 (1995).
[10] S. Amari, A. Cichocki, and H. H. Yang, in *Advances in neural information processing systems*, edited by D. Touretzky, M. Mozer, and M. Hasselmo (Cambridge MA, 1996), vol. 8, pp. 757–763.
[11] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Cell **138**, 774 (2009).
[12] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, Nucl. Acids Res. **28**, 33 (2000).

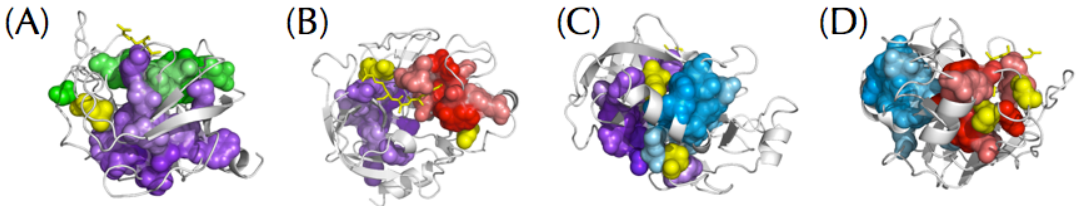

FIG. S 1: Intersections between $\mathcal{S}_k$ – In the main text, sectors are defined by $\mathcal{S}'_k = \{i : V_i^{(k)} > \epsilon, V_i^{(\ell)} < \epsilon,\ \ell \neq k\}$ ($k = 1, \ldots, k_{\rm top} = 4$, $\epsilon = 0.1$). Here, we represent in yellow the few positions that belong to the intersections $\mathcal{S}_k \cap \mathcal{S}_\ell$ for $k \neq \ell$, with $\mathcal{S}_k = \{i : V_i^{(k)} > \epsilon\}$. There are 7 such positions, all of which are at the structural periphery of either $\mathcal{S}'_k$ or $\mathcal{S}'_\ell$, including 4 that are at the structural interface between the two. (A position from the green sector appears as disconnected in (A), but the position that could connect it to the rest of the green sector is actually not included in the alignment).

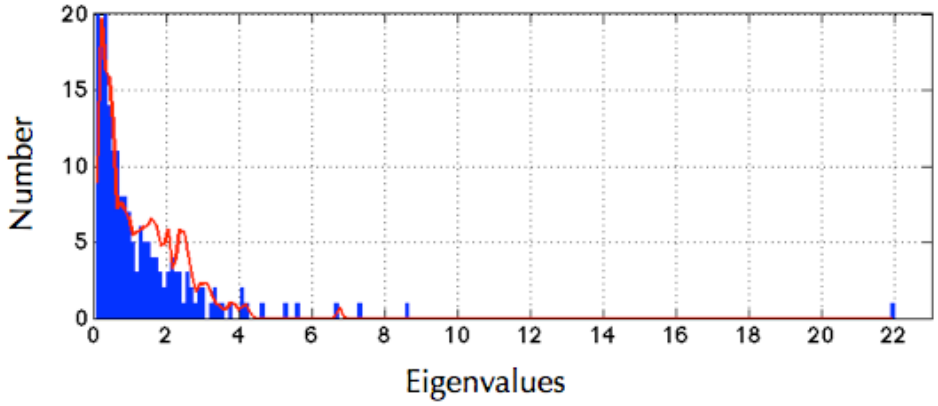

FIG. S 2: Spectrum of the SCA matrix $\mathcal{C}_{ij}$ – In blue, histogram of the L eigenvalues of the matrix $\mathcal{C}_{ij}$ (truncated to 20 along the y-axis). In red, average spectrum over 100 randomized alignments, where the amino acids are drawn independently at each position i according to the frequencies f_i^a. This shows that between 3 and 7 eigenvalues may considered as significant.

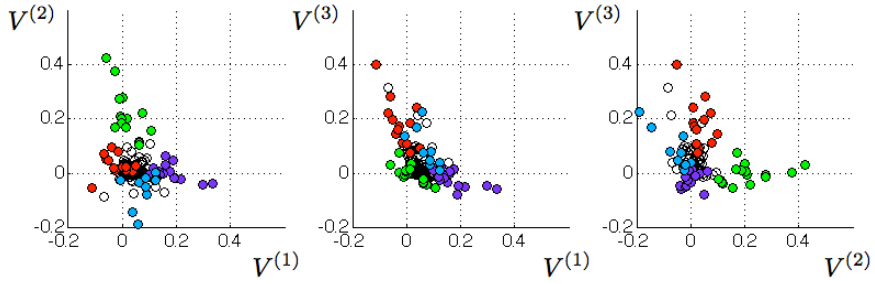

FIG. S 3: Independent components from rotation by ICA of the top $k_{\rm top} = 3$ eigenvectors of $\mathcal{C}_{ij}$ – This figure is the counterpart of Fig. 1, for $k_{\rm top} = 3$ instead of $k_{\rm top} = 4$, and with positions colored as in Fig. 1. The same green and red sectors are defined along $V^{(1)}$ and $V^{(2)}$, but the red and cyan sectors appear together along $V^{(3)}$.

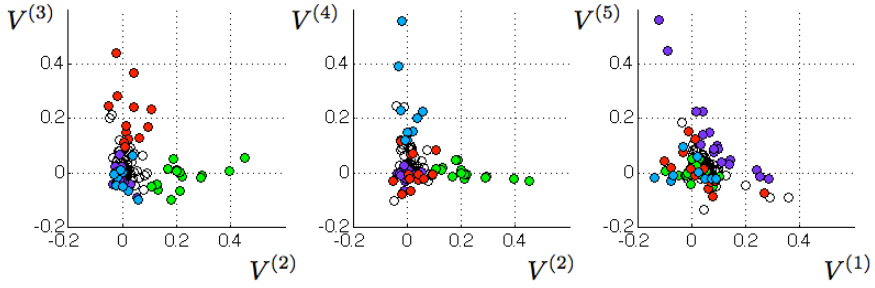

FIG. S 4: Independent components from rotation by ICA of the top $k_{\rm top} = 5$ eigenvectors of $\mathcal{C}_{ij}$ – This figure is the counterpart of Fig. 1, for $k_{\rm top} = 5$ instead of $k_{\rm top} = 4$, and with positions colored as in Fig. 1. The definitions of the green, red and cyan sectors along $V^{(2)}, V^{(3)}, V^{(4)}$ are consistent with their definition with $k_{\rm top} = 4$, while the new component $V^{(5)}$ decomposes the purple sector in two subsets, whose interpretation is given in Fig. S5.

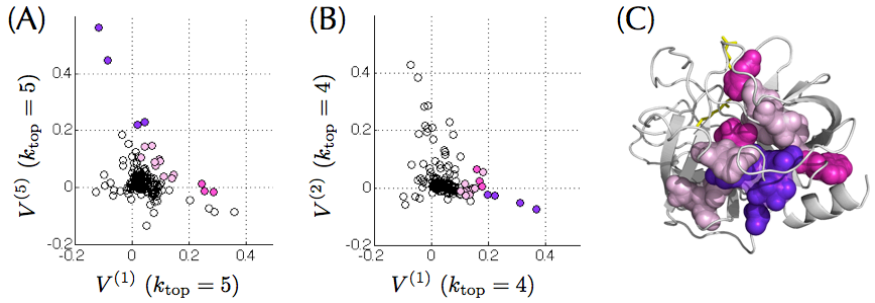

FIG. S 5: Interpretation of the decomposition of the purple sector when considering $k_{\rm top} = 5$ – **(A)** Same as the last graph of Fig. S4, except that the positions from the purple sector are indicated with three different colors, depending on whether they satisfy $V_i^{(1)} > 0.2$ (pink), $V_i^{(5)} > 0.2$ (purple) or neither (light pink). **(B)** Using the same coloring scheme, but now for the projection along components obtained with $k_{\rm top} = 4$ as in Fig. 1, shows that the partition of the purple sector corresponds to a partition between a core, defined as the positions with highest contribution to $V^{(1)}$, and other positions. **(C)** Location of the positions on the three-dimensional structure, showing that the core is structurally connected with the other positions around.

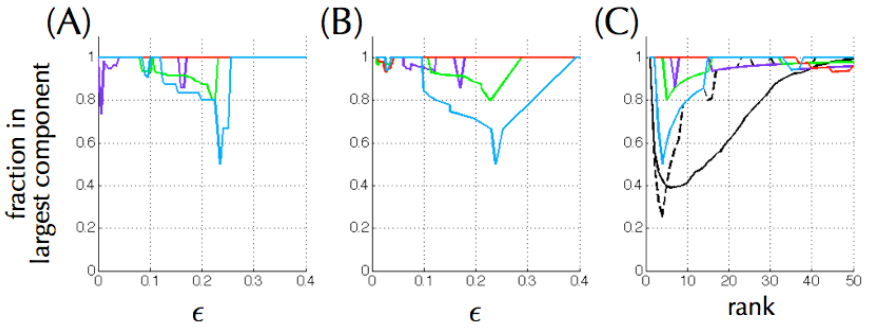

FIG. S 6: Connectivity of the sectors for varying values of ϵ – **(A)** Fraction of the positions in the largest structurally connected subset of a sector for varying values of ϵ and for sectors defined as in the main text by $\mathcal{S}'_k = \{i : V_i^{(k)} > \epsilon, V_i^{(\ell)} < \epsilon,\ \ell \neq k\}$ ($k = 1, \ldots, k_{\rm top} = 4$). The colors correspond to those of Fig. 1: magenta for $k = 1$, green for $k = 2$, red for $k = 3$ and cyan for $k = 4$. **(B)** Same as (A) but not excluding intersections, which are significant for small ϵ, i.e., taking $\mathcal{S}_k = \{i : V_i^{(k)} > \epsilon\}$. **(C)** To assess the significance of (A) and (B), we consider also two other groups of ordered positions: randomly ranked positions and positions ranked by degree of conservation, measured by the relative entropy $D_i = \sum_{a=0}^{20} f_i^a \ln f_i^a/q^a$ with $f_i^0 = 1 - \sum_{a=1}^{20} f_i^a$ for the frequency of gaps at position i and $q^a = \sum_i f_i^a/L$ for the background frequencies. The results are presented here as a function of the size of the groups, where positions are added according to their rank (given by $V_i^{(k)}$ for the sectors). Compared to randomly ranked positions (full black line), sectors are clearly significantly more connected. They are also more connected than positions ranked by conservation (dashed black line), with the possible exception of the cyan sector.

Purple sector	Green sector	Red sector	Cyan sector
29	57*	228*	237*
122	195*	189*	238
46*	197*	215*	234
120	55*	183*	91
28	19*	226*	231
51	102*	216*	103
40	43*	213*	229*
30	194*	227*	101
104*	193	184	123*
27	214*	191*	199
31	94	192*	
136*	142*	172*	
52*	42*	138	
201*	58*		
32	33*		
68*			
200			

TAB. S I: Composition of the sectors defined in Fig. 1, ranked by the corresponding value of $V_i^{(k)}$ – A star indicates for the purple and cyan sectors that the position belongs to the blue sector defined in Ref. [11], and for the green and red sectors that it belongs respectively to the green and red sector of Ref. [11]. Note that 19 positions included in the alignment of Ref. [11] are not represented in the present alignment, including 1 of the 20 positions of the red sector defined in Ref. [11], 0 of the 22 positions of its green sector and 4 of the 25 positions of its blue sector.

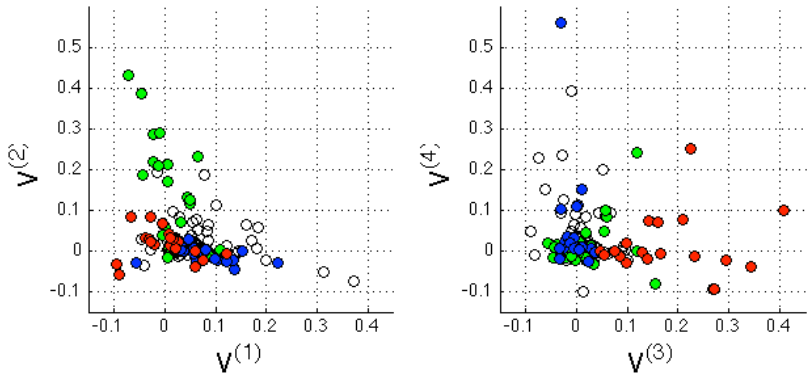

FIG. S 7: Comparison with the sectors defined in Ref. [11] – The graphs are identical to those of Fig. 1(A), except that the positions are colored according to the definition of the 3 sectors of Ref. [11]. This shows as in Tab. SI that essentially the same green and red sectors are identified, while the purple and cyan sectors have a small overlap with the previously defined blue sector.

contact	pair	secton(s)	dist. (Å)
1	42 - 58	2	2
2	136 - 201	5	2
3	57 - 195	3	2.7
4	191 - 220	6	2.2
5	168 - 182	4	2
6	32 - 40	7	2.8
9	189 - 226	8	3.3
11	59 - 104	11	3.9
16	29 - 122	15	6.7
17	26 - 157	10	4.9
20	44 - 52	9	4.3
21	142 - 194	12	2.7
25	30 - 139	13	2.7
36	51 - 105	16	3.8
37	46 - 112	14	4
38	27 - 157	10	3.8
40	161 - 184	19	3.1
44	146 - 221A	21	3.6
54	68 - 112	14	3.8
58	51 - 107	16	3.6
59	53 - 209	17	3.5
64	57 - 102	3	2.7
71	21 - 156	24	3.3
75	31 - 44	9	3.1
76	86 - 107	22	2.7
77	45 - 209	17	3.8
78	100 - 179	25	3
80	158 - 188A	26	3.5
83	138 - 190	18	4.7
84	190 - 213	18	3.7
99	189 - 228	44	3.9
104	138 - 213	18	4.2
106	102 - 195	3	6.1
112	103 - 234	20	4.1
114	22 - 157	-	2
119	31 - 68	84	3.8
123	31 - 52	9	4
125	30 - 155	-	4
126	180 - 215	23	3.8
127	215 - 227	23	2.8

contact	pair	secton(s)	dist. (Å)
128	27 - 139	-	3.6
129	189 - 221	44	3
130	81 - 118	28	3.8
134	136 - 162	-	3.3
135	136 - 199	-	3.8
137	165 - 230	32	5
141	50 - 107	22	3.3
142	183 - 189	-	4.7
144	29 - 200	-	3.5
145	46 - 68	14	3.9
146	183 - 199	27	4.3
147	138 - 158	26	2.8
148	136 - 210	-	5.4
153	45 - 121	17 31	4.1
154	189 - 216	-	6.2
155	121 - 209	17	4
156	195 - 214	-	3.4
159	46 - 52	-	3.8
161	100 - 177	59	2.6
164	67 - 82	35	3.5
167	103 - 229	20	4.2
169	181 - 199	27	6.6
171	139 - 200	-	6.7
173	114 - 120	33	3.9
174	55 - 195	-	4.7
175	85 - 106	29	3.4
178	91 - 101	95	3.6
182	101 - 234	95	2.6
184	84 - 109	38	2.9
189	50 - 111	37	3.6
192	102 - 214	41	2.5
199	195 - 213	-	3.7
200	142 - 193	-	3.7
201	27 - 137	39	4.4
203	87 - 107	85	2.8
209	46 - 108	-	4.3
211	195 - 216	-	7.8
214	161 - 184A	-	3.8
217	179 - 233	25	3
218	182 - 191	-	13

contact	pair	secton(s)	dist. (Å)
223	138 - 160	98	4.1
227	100 - 180	-	3.8
231	46 - 114	-	3.6
233	124 - 235	34	3.8
235	72 - 78	40	8
236	30 - 198	13	3.3
240	183 - 226	-	3.2
241	27 - 122	-	11
243	83 - 112	-	4
244	57 - 213	-	4.5
245	139 - 157	-	4.2
250	190 - 216	45	7.1
251	68 - 118	-	4
256	57 - 193	-	7.7
261	144 - 152	43	3.8
264	140 - 194	-	3.3
266	45 - 212	30	6.3
268	101 - 179	-	3.3
270	216 - 227	45	4.7
273	49 - 111	37	3.2
274	163 - 225	42	3.7
280	43 - 57	-	8.3
285	139 - 198	13	3.6
286	55 - 197	36	5.7
289	219 - 226	-	6.6
292	21 - 154	89	3.8
293	86 - 109	-	3.9
296	30 - 43	-	4
298	179 - 234	-	3.2
300	122 - 189	-	21
301	43 - 142	-	7.8
309	83 - 109	-	3.8
311	90 - 104	-	3.9
314	199 - 228	-	3.3
321	160 - 228	98	6.8
328	29 - 124	-	12
333	180 - 189	-	13
334	135 - 161	-	3.2
337	121 - 200	31 118	3.9
338	180 - 227	23	3.8

TAB. S II: Top 120 non-trivial contacts and their associated secton – The pairs are ordered by decreasing values of $\mathcal{D}_{ij}$, with their rank indicated in the first column. Only pairs of positions separated by $> \Delta = 5$ amino acids along the sequence are considered, and many pairs are therefore not included (red curve in Fig. S8). The second column indicates the rank of the secton where the pair is found and the third the distance in angstroms between the positions in the three-dimensional structure. Pairs are considered in contact when this distance is < 8 Å, and false positive are indicated in red. Here, none of these false positive are in a secton, but many true positive are also not in a secton.

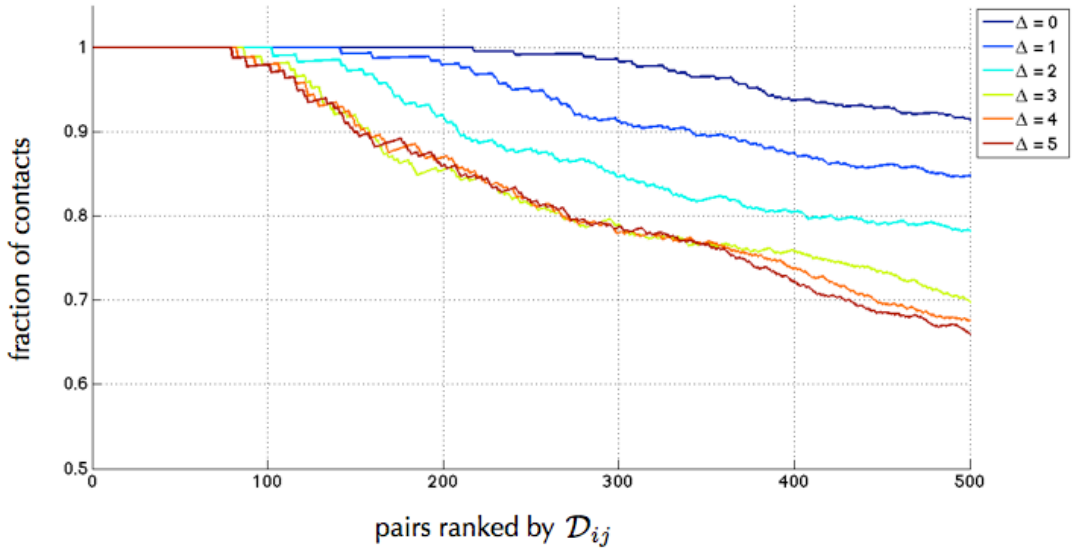

FIG. S 8: Fraction of pairs of positions predicted to be in contact from the top values of the matrix of direct information $\mathcal{D}_{ij}$ – Pairs ij of positions are ordered by decreasing values of $\mathcal{D}_{ij}$, and the fraction of top n pairs to be in structural contact (distance < 8 Å) is considered as a function of n. Different curves correspond to different values of Δ, where pairs separated by $\leq \Delta$ amino acids along the sequence are ignored.

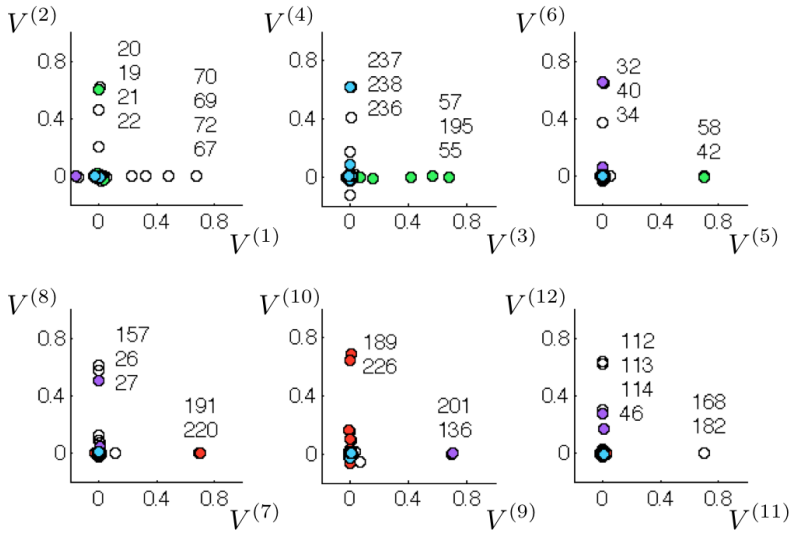

FIG. S 9: Top 24 sectons when not truncating $\mathcal{D}_{ij}$ to $\tilde{\mathcal{D}}_{ij}$ (except for the strict diagonal, $\mathcal{D}_{ij} = 0$) – Out of the 24 sectons displayed here, 5 consist exclusively of consecutive positions and may be attributed to consecutive gaps.

Sectons	Positions	Sectons	Positions	Sectons	Positions	Sectons	Positions
*1	224 217	*31	121 45 200	*61	175	*91	96
*2	58 42	*32	165 230 176	62	130 123 166	*92	23
*3	57 195 102	*33	114 120 48	*63	28 119	*93	171
*4	168 182	*34	235 124 123	*64	110 83	*94	167
*5	201 136	*35	67 82 34	*65	117 25	*95	91 101 56 234
*6	191 220	36	197 55 43 19	66	98 166	*96	115
*7	32 40	*37	111 50 49 83	*67	151 143	*97	88
*8	226 189	*38	109 84	*68	95	*98	160 228 138
*9	44 52 31	*39	27 137	*69	74 153	99	19 25 33
*10	157 26 27	*40	72 78	*70	47 238 53	*100	135 159
*11	59 104	*41	214 102 229	*71	148	*101	60
*12	194 142	*42	163 225	*72	39	*102	24 70 73
*13	30 139 198	*43	144 152	73	37 124	*103	203
*14	112 46 68	*44	221 228 189	*74	22 155	*104	211 228
*15	29 122	*45	216 227 190	*75	204	105	48 125
*16	51 105 107	*46	164 132	*76	193 55 43	*106	236
*17	209 53 45 121	47	210 198 162 124 225	*77	75 67	*107	61
*18	190 213 138	48	196 140	*78	173	*108	97
*19	184 161	*49	77 70	*79	89	*109	170
*20	103 234 229	*50	133 162	*80	54 33	*110	63
*21	221A 146	*51	192 143 219	*81	33 41 66	111	49 113 233
*22	107 86 50	*52	108 66 83	82	188 181 145 221 169	*112	129
*23	215 227 180	*53	94 56	*83	64 39	113	159 137 119 23 176
*24	156 21	*54	176 169	*84	31 68 66	*114	147 148
*25	179 100 233	*55	231 238 123	*85	87 107	*115	178
*26	158 188A 138	*56	93 101	*86	116	116	34 153
*27	199 181 183	57	128 232 22	*87	184A 221	117	134 113 62 25
*28	118 81 69	*58	92 237	*88	174	*118	200 121
*29	85 106 66	*59	177 100	*89	154 21	*119	141 69
*30	212 55 45 229	*60	222 187 172	*90	76	120	125 234

TAB. S III: Composition of sectons – Sectons are defined here with $k_{\text{top}} = 120$ and $\epsilon = 0.2$. A star indicates that the secton is structurally connected. Up to rank 35, all sectons are thus connected. Note that at large rank, some sectons consist of a single position and are therefore trivially connected; see also Fig. S11. As indicated in Fig. S11, structurally disconnected sectons often become connected when considering a more stringent threshold ϵ; this is for instance the case for secton 36 when excluding its last position, 19.

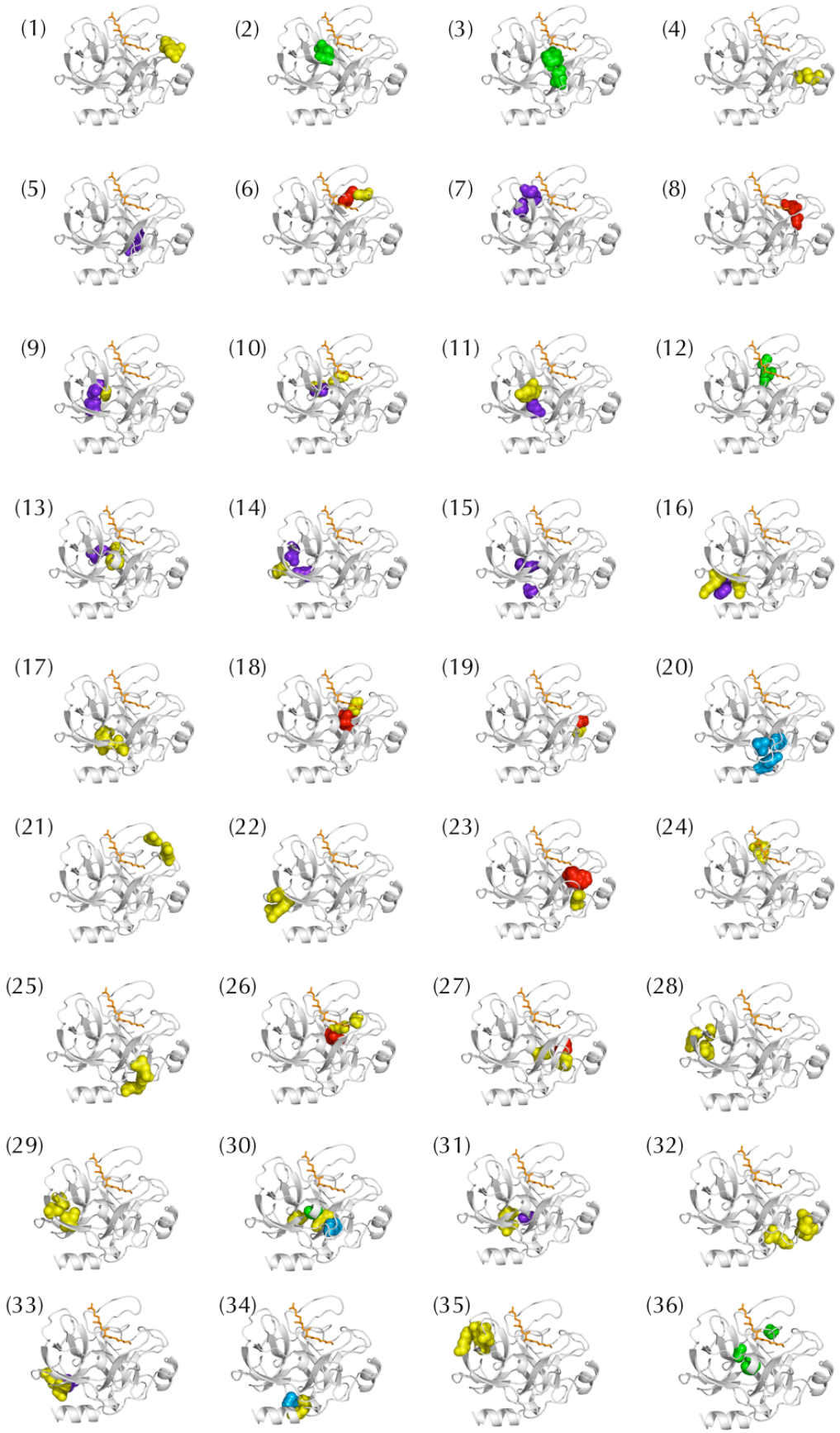

FIG. S 10: Structural representation of the top 36 sectons – Representation on the three-dimensional structure of rat trypsin of the top 36 sectons defined in Tab. SIII. The colors refer to the sectors, with yellow for non-sector positions (as in Fig. 3, where the first 8 sectons are represented). The 36th secton is the first not to be completely structurally connected, with one position standing apart.

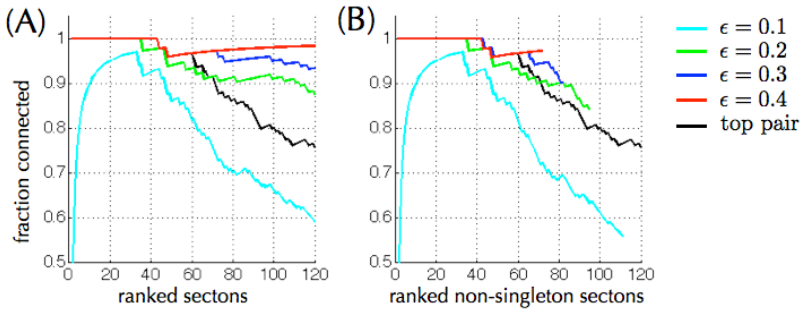

FIG. S 11: Connectivity of sectons – Sectons are here defined for $k_{\text{top}} = 120$ and different values of ϵ. A secton is considered as structurally connected if all the positions that it contains are in contact in the three-dimensional structure, either directly or indirectly. **(A)** Fraction of the top sectons to be structurally connected as function of the rank up to which sectons are considered. For comparison, the black curve corresponds to the top 2 positions contributing to $V^{(k)}$. **(B)** Same as (A), but excluding the sectons that contain a single position and are therefore trivially connected. These figures show that by considering a larger, more stringent threshold ϵ than in Tab. SIII (where $\epsilon = 0.2$) more (but smaller) sectons are found to be connected.

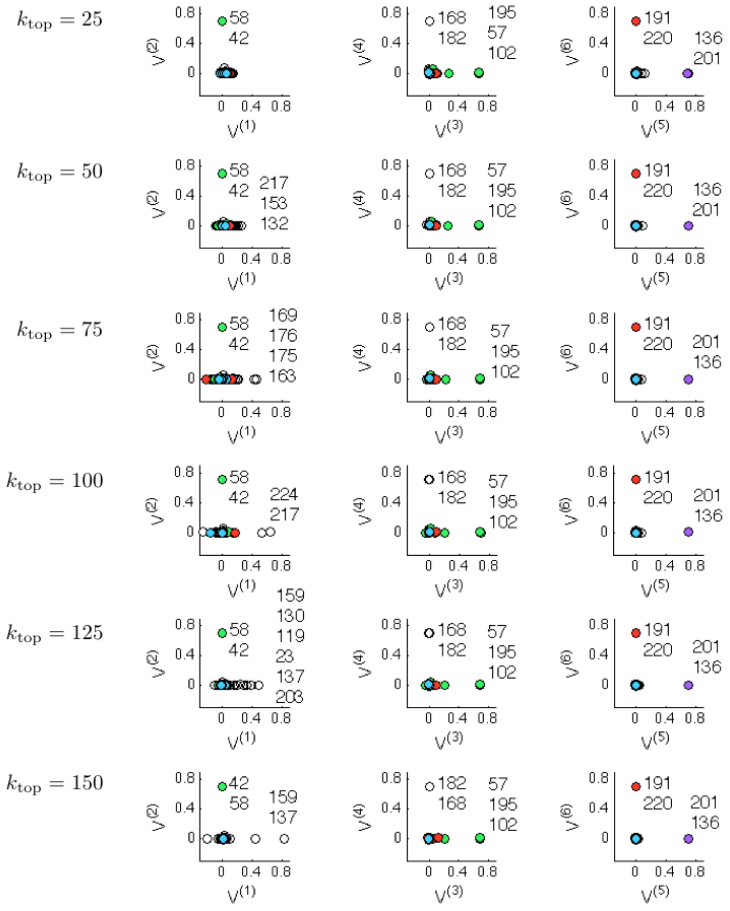

FIG. S 12: Top independent components of $\tilde{\mathcal{D}}_{ij}$ for varying values of k_{top} – This figure generalizes Fig. 3, where $k_{\text{top}} = 120$, to $k_{\text{top}} = 25, 50, 75, 100, 125, 150$. With the noticeable exception of the first component $V^{(1)}$, the same components $V^{(k)}$, and therefore the same sectons, are recovered for different values of k_{top}. The instability of $V^{(1)}$ suggests that the first secton may better be discarded (although, for $k_{\text{top}} = 100$ and $k_{\text{top}} = 150$ at least, it does appear here to convey meaningful information, as the pairs 217-224 and 137-159 are indeed in contact).

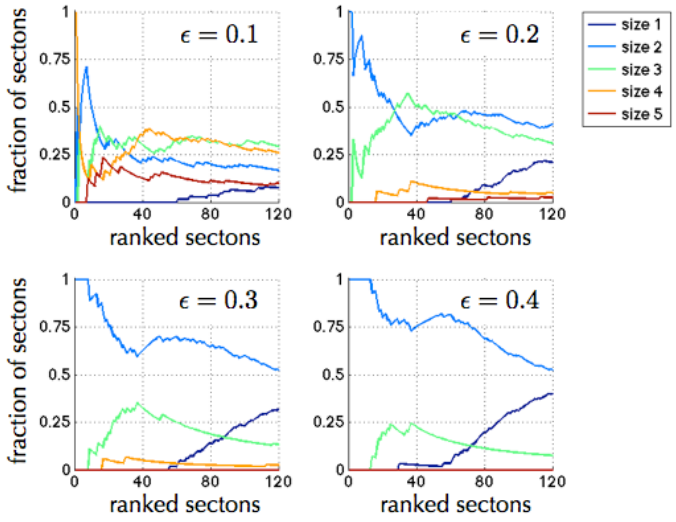

FIG. S 13: Size of sectons – Distribution of the sizes of sectons, for different values of ϵ. The larger the cut-off ϵ is, the smaller are the sectons.

Subgraph	Positions	Secton (if applicable)
1	21 156	24
2	32 40	7
3	42 58	2
4	67 82	35^-
5	81 118	28^-
6	84 109	38
7	85 106	29^-
8	59 104	11
9	114 120	33^-
10	146 221A	21
11	165 230	32^-
12	168 182	4
13	191 220	6
14	142 193 194	12^+
15	161 184 184A	19^+
16	180 215 227	23
17	45 53 121 209	17
18	100 177 179 233	
19	91 101 103 229 234	
20	31 44 46 52 68 108 112	
21	50 51 86 87 105 107 111	
22	22 26 27 29 30 122 137 139 155 157 200	
23	55 57 102 136 138 158 162 181 183 188A	
	189 190 195 199 201 210 213 214 216 221 226 228	

TAB. S IV: Decomposition into connected subgraphs of the contact graph defined from the top pairs of $\tilde{\mathcal{D}}_{ij}$ – The top 79 pairs of $\tilde{\mathcal{D}}_{ij}$, which are all structural contacts, define a contact graph where two pairs are linked if in physical contact (distance < 8 Å). We list here the disjoint connected subgraphs, ranked by size, of this graph, which we compare to sectons. Of the 13 subgraphs of size 2, 7 correspond to sectons of size 2 and 5 to sectons of size 3 (indicated by $^-$). Of the 5 subgraphs of size 3 and 4, 3 correspond to sectons of same size and 2 to sectons of smaller size 3 (indicated by $^+$). Larger subgraphs are also found that generally consist of multiple sectons. This table shows that the sectons are not equivalent to the connected subgraphs of the contact graph defined from the top pairs of $\tilde{\mathcal{D}}_{ij}$.

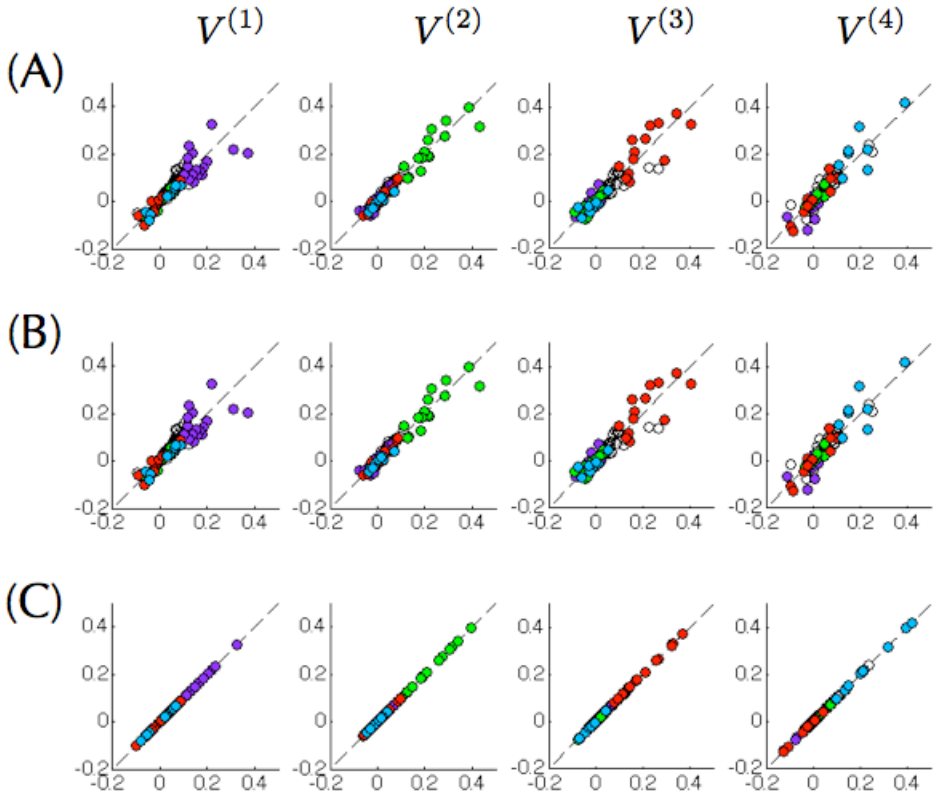

FIG. S 14: – Comparison of the top $k_{\text{top}} = 4$ independent components of SCA matrices defined from the covariance matrix C, $\bar{C}$ and $\bar{C}^{+}$ – (**A**) $V^{(k)}$ from a regularized covariance matrix $\bar{C}$ ($\mu = 1/2$) against $V^{(k)}$ from the original covariance matrix C ($\mu = 0$). Positions are colored as in Fig. 1. (**B**) $V^{(k)}$ from a regularized covariance matrix $\bar{C}^{+}$ truncated from its top k^* modes ($\mu = 1/2$, $k^* = 100$) against $V^{(k)}$ from the original covariance matrix C ($\mu = 0$, $k^* = 0$). (**C**) $V^{(k)}$ from a regularized covariance matrix $\bar{C}^{+}$ truncated from its top k^* modes ($\mu = 1/2$, $k^* = 100$) against $V^{(k)}$ from the untruncated covariance matrix C ($\mu = 1/2$, $k^* = 0$). The regularization has a small incidence on the definition of sectors, while the truncation from the top modes has none.

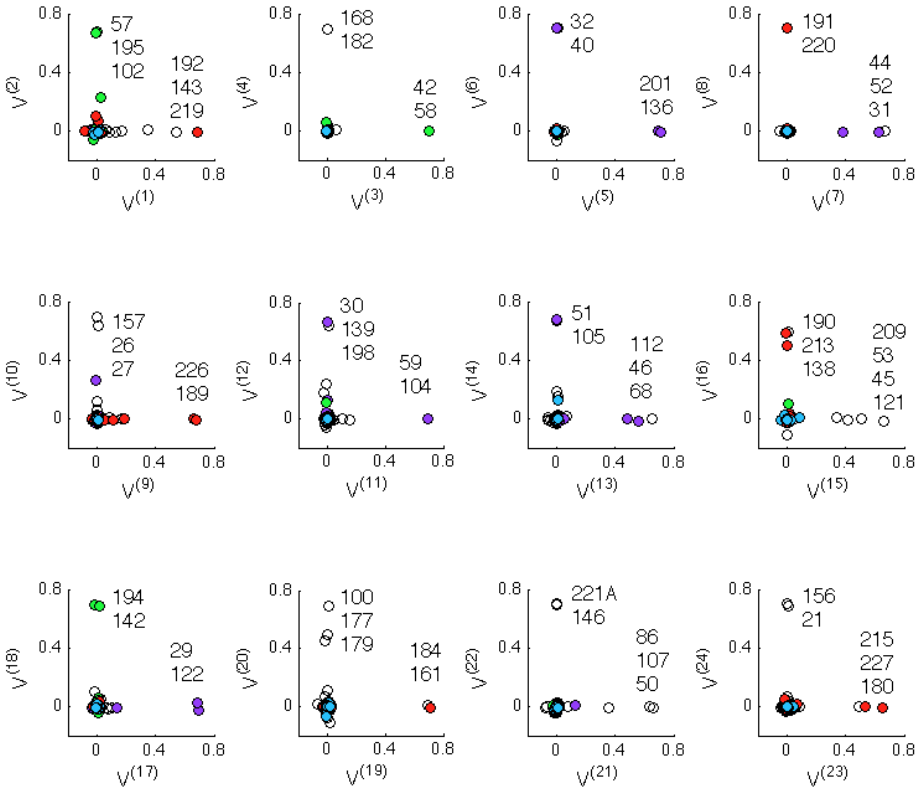

FIG. S 15: Sectons from the bottom modes of $\bar{C}$ – We show here the top 24 sectons derived from $\tilde{\mathcal{D}}_{ij}[J^-]$, with $k^* = 100$. A comparison with Fig. 2 shows that essentially the same sectons are defined by this matrix and by $\tilde{\mathcal{D}}_{ij}[J]$. The exception of the first secton is to be related related to the unstability reported in Fig. S12.

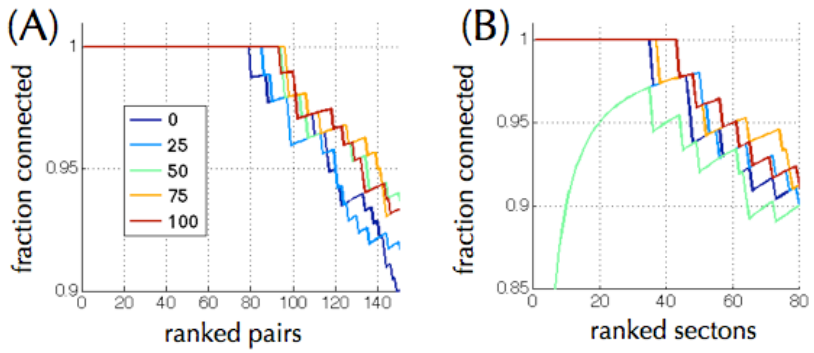

FIG. S 16: Top pairs and sectons from covariance matrices $\bar{C}$ cleaned from their top modes – As in the main text, we consider $\bar{C} = \sum_k |k\rangle\lambda_k\langle k|$, the spectral decomposition of $\bar{C}$ in the bra-ket notation, with ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_L$. The truncated matrix of direct information $\tilde{\mathcal{D}}_{ij}$ is defined from $J^- = -\sum_{k>k^*} |k\rangle\lambda_k^{-1}\langle k|$ for different values of k^*, $k^* = 0, 25, 50, 75, 100$, corresponding to the different curves. Panel (A) for $k^* = 0$ (dark blue) corresponds to the results of Fig. S8 for $\Delta = 5$ and panel (B) for $k^* = 0$ to Fig. S11(A) for $\epsilon = 0.2$. Truncating the top $k^* = 100$ modes of $\bar{C}$ thus leads to an increase of $\sim 20\%$ in the number of top pairs and top sectons in physical contact. In (B), $k^* = 50$ stands apart as a consequence of a disconnected first secton, which, following the observation made in Fig. S12, could have been ignored (the first 2 positions of this secton are actually connected).

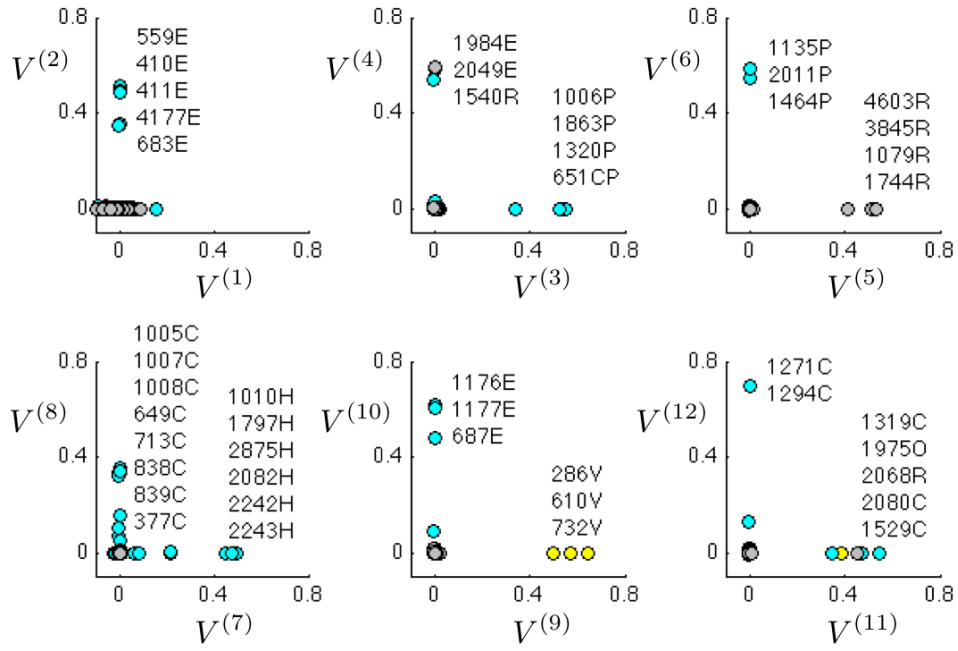

FIG. S 17: Top genomic sectons in bacteria – As Fig. 3, but for the co-occurrence of orthologous genes in bacterial genomes. Each dot represents a cluster of orthologous genes (COG), with a color associated with its functional class: cyan for metabolism, yellow for cellular processes, magenta for information processing, and gray for poorly characterized genes [12]. The sectons often comprise genes from a common functional subclass, which is indicated by the last letter labeling the COGs (see Tab. SV for further genomic sectons).

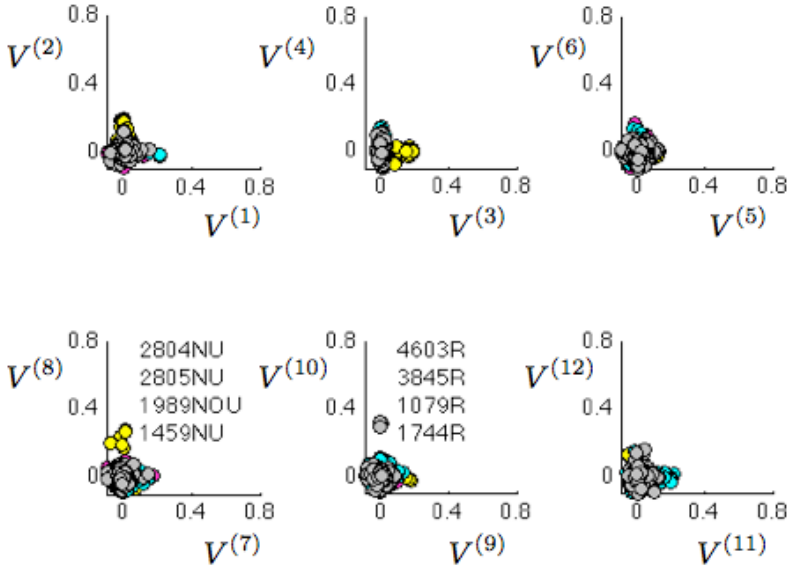

FIG. S 18: Secton analysis of C_{ij} – The same approach used to define sectons from $\mathcal{D}_{ij}$ is applied to C_{ij}. In contrast to Fig. S17, the components are here not localized, thus not defining 'sectons' (truncating the diagonal of C_{ij} does not alter this conclusion).

secton	COGs	secton	COGs
1		51	1366T 2172T 2208TK
2	559E 410E 411E 4177E 683E	52	1108P 1121P 803P
3	1006P 1863P 1320P 651CP	53	5405O 1220O
4	1984E 2049E 1540R	54	1682GM 1134GM
5	4603R 3845R 1079R 1744R	55	280C 282C
6	1135P 2011P 1464P	56	2884D 2177D
7	1010H 1797H 2875H 2082H 2242H 2243H	57	2224C 2225C
8	1005C 1007C 1008C 649C 713C 838C 839C 377C	58	1126E 765E 834ET
9	286V 610V 732V	59	106E 107E 118E 40E 131E 139E 141E
10	1176E 1177E 687E	60	1663M 1043M 1212M 2877M [...]
11	1319C 1975O 2068R 2080C 1529C	61	1392P 306P
12	1271C 1294C	62	588G 696G
13	163H 43H	63	1354S 1386K
14	1788I 2057I	64	168P 569P
15	1346M 1380R	65	1290C 2010C 723C
16	1653G 3839G 395G 1175G	66	45C 74C
17	1732M 1125E	67	2801L 2963L
18	1116P 600P 715P	68	245I 1154HI 1211I 1947I 743I 761IM 821I
19	1203R 1518L	69	175EH 2895P 7H 155P
20	1638G 1593G	70	1122P 619P
21	1640G 58G 296G 297G 1523G 448G	71	1183I 688I
22	1088M 1091M 1209M 1898M	72	1080G 1762GT 1925G
23	1129G 1172G 1879G	73	132H 502H 156H 161H
24	2894D 850D 851D	74	751J 752J
25	396O 719O	75	1828F 47F
26	3451U 3505U	76	2009C 1485R 2142C
27	2332O 2386O 1138O	77	263E 14E
28	1270H 2038H 2087H 1492H 368H	78	4799I 511I 777I 825I
29	1677NU 1684NU 1766NU 1157NU [...]	79	1180O 1328F
30	22C 1071C 508C	80	1013C 674C
31	1228Q 2986E 2987E	81	113H 181H 1H 373H
32	1120PH 609P 614P	82	3275T 3279KT
33	1622C 109O 1845C 843C 1612O	83	413H 414H 853H
34	3288C 1282C	84	547E 133E 134E 135E 147EH 512EH 159E
35	1918P 370P	85	1704S 1512R
36	2804NU 1989NOU 1450NU 1459NU	86	554C 578C
37	117H 1985H	87	602O 603R 720H
38	2896H 303H 746H 521H 314H 315H	88	1077D 1792M 1426S
39	4962U 4965U	89	4149P 725P
40	241E 279G 2870M 859M	90	3383R 437C 1526C
41	1003E 404E 509E	91	1020Q 2091H 736I
42	2025C 2086C	92	1706N 1261NO 2063N
43	422H 2022H 2104H	93	55C 56C 355C 356C 711C 712C 224C
44	1173EP 601EP 747E 444EP	94	473CE 119E 129EG 59EH 65E 66E 440E
45	226P 1117P 573P 581P	95	1837R 1847R
46	379H 29H 157H	96	363G 364G
47	1352NT 643NT 2201NT 835NT 840NT	97	2805NU 1989NOU 4972NU
48	1178P 1840P	98	548E 1364E 4992E 2E
49	1034C 1894C 1905C	99	419L 420L
50	1127Q 1463Q 767Q	100	455D 1191K 1317NU 1345N 1419N 1516NUO 3144N

TAB. S V: Content of the top 100 genomic sectons in terms of COGs – The full content of secton 29 is {1677NU 1684NU 1766NU 1157NU 1815N 1843N 1868N 1256N 1291N 1298NU 1338NU 1344N 1987NU 1377NU 1536N 1558N} and the full content of secton 60 is {1663M 1043M 1212M 2877M 763M 774M 1519M 794M 1560M}.

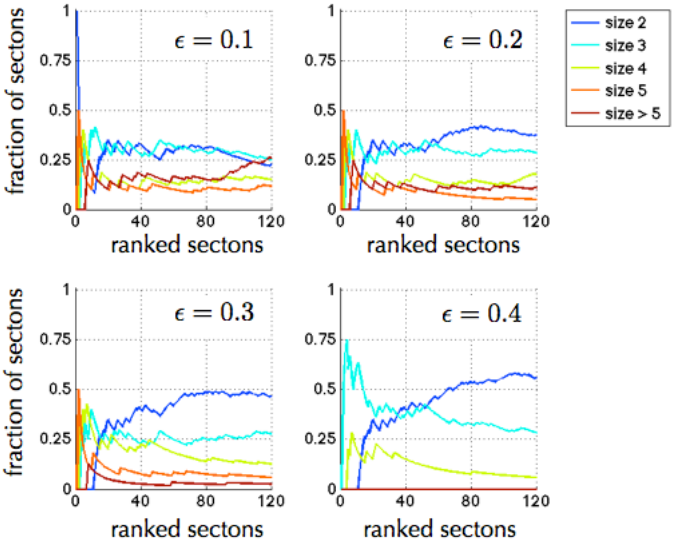

FIG. S 19: Size of genomic sectons – Distribution of the sizes of sectons, for different values of ϵ. The larger the cut-off ϵ is, the smaller are the sectons (sectons of size 1 are in negligible number).

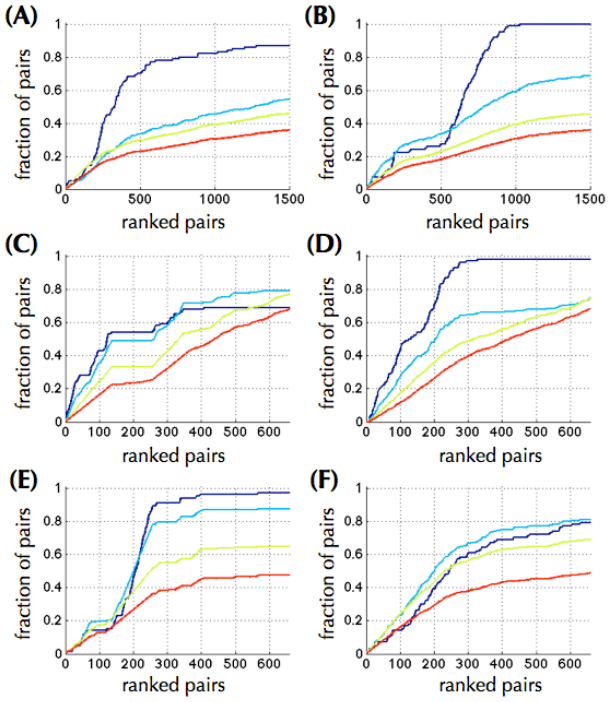

FIG. S 20: Comparison of top pairs for C_{ij}, $\mathcal{D}_{ij}$ and sectons – **(A)** To compare the top pairs for C_{ij} and $\mathcal{D}_{ij}$, we represent the fraction of top pairs for C_{ij}, up to the rank indicated along the x-axis, to be included in the top 100 (blue curve), 500 (cyan), 1000 (green) or 1500 (red) top pairs for $\mathcal{D}_{ij}$. This shows that $\sim 40\%$ of the top n pairs for C_{ij} are in the top n pairs for C_{ij}. **(B)** Same as (A) but exchanging the roles of C_{ij} and $\mathcal{D}_{ij}$. **(C)** To compare with the content of sectons, we listed the pairs contained in sectons, ranked them by rank of sectons (with arbitrary ranking within sectons) and represented the fraction of these top pairs, up to the rank indicated along the x-axis, to be included in the top 100 (blue curve), 200 (cyan), 400 (green) or 600 (red) top pairs for $\mathcal{D}_{ij}$. **(D)** Same as (C) but exchanging the roles of top pairs for sectons and top pairs for $\mathcal{D}_{ij}$. **(E)** Same as (C) but for C_{ij} instead of $\mathcal{D}_{ij}$. **(F)** Same as (D) but for C_{ij} instead of $\mathcal{D}_{ij}$. Overall, the last 4 panels show an overlap of $\sim 60\%$ between pairs in top sectons and top pairs for $\mathcal{D}_{ij}$ and $\sim 40\%$ between pairs in top sectons and top pairs for C_{ij}.

Secton	COG	Annotation
2	559E	Branched-chain amino acid ABC-type transport system, permease components
	410E	ABC-type branched-chain amino acid transport systems, ATPase component
	411E	ABC-type branched-chain amino acid transport systems, ATPase component
	4177E	ABC-type branched-chain amino acid transport system, permease component
	683E	ABC-type branched-chain amino acid transport systems, periplasmic component
3	1006P	Multisubunit Na+/H+ antiporter, MnhC subunit
	1863P	Multisubunit Na+/H+ antiporter, MnhE subunit
	1320P	Multisubunit Na+/H+ antiporter, MnhG subunit
	651CP	Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MnhD subunit
4	1984E	Allophanate hydrolase subunit 2
	2049E	Allophanate hydrolase subunit 1
	1540R	Uncharacterized proteins, homologs of lactam utilization protein B
5	4603R	ABC-type uncharacterized transport system, permease component
	3845R	ABC-type uncharacterized transport systems, ATPase components
	1079R	Uncharacterized ABC-type transport system, permease component
	1744R	Uncharacterized ABC-type transport system, periplasmic component/surface lipoprotein
6	1135P	ABC-type metal ion transport system, ATPase component
	2011P	ABC-type metal ion transport system, permease component
	1464P	ABC-type metal ion transport system, periplasmic component/surface antigen
7	1010H	Precorrin-3B methylase
	1797H	Cobyrinic acid a,c-diamide synthase
	2875H	Precorrin-4 methylase
	2082H	Precorrin isomerase
	2242H	Precorrin-6B methylase 2
	2243H	Precorrin-2 methylase
8	1005C	NADH:ubiquinone oxidoreductase subunit 1 (chain H)
	1007C	NADH:ubiquinone oxidoreductase subunit 2 (chain N)
	1008C	NADH:ubiquinone oxidoreductase subunit 4 (chain M)
	649C	NADH:ubiquinone oxidoreductase 49 kD subunit 7
	713C	NADH:ubiquinone oxidoreductase subunit 11 or 4L (chain K)
	838C	NADH:ubiquinone oxidoreductase subunit 3 (chain A)
	839C	NADH:ubiquinone oxidoreductase subunit 6 (chain J)
	377C	NADH:ubiquinone oxidoreductase 20 kD subunit and related Fe-S oxidoreductases
9	286V	Type I restriction-modification system methyltransferase subunit
	610V	Type I site-specific restriction-modification system, R (restriction) subunit and related helicases
	732V	Restriction endonuclease S subunits
10	1176E	ABC-type spermidine/putrescine transport system, permease component I
	1177E	ABC-type spermidine/putrescine transport system, permease component II
	687E	Spermidine/putrescine-binding periplasmic protein
11	1319C	Aerobic-type carbon monoxide dehydrogenase, middle subunit CoxM/CutM homologs
	1975O	Xanthine and CO dehydrogenases maturation factor, XdhC/CoxF family
	2068R	Uncharacterized MobA-related protein
	2080C	Aerobic-type carbon monoxide dehydrogenase, small subunit CoxS/CutS homologs
	1529C	Aerobic-type carbon monoxide dehydrogenase, large subunit CoxL/CutL homologs
12	1271C	Cytochrome bd-type quinol oxidase, subunit 1
	1294C	Cytochrome bd-type quinol oxidase, subunit 2

TAB. S VI: Annotations for the top genomic sectons – Note that there is no secton along the first component, as seen in Fig. S17. (Sections 13 to 24 are presented on the next page). This shows that the definition of sectons is consistent with our current knowledge of gene functions with, in particular, many of the top sectons involving different subunits of a same protein complex.

11	1319C	Aerobic-type carbon monoxide dehydrogenase, middle subunit CoxM/CutM homologs
	1975O	Xanthine and CO dehydrogenases maturation factor, XdhC/CoxF family
	2068R	Uncharacterized MobA-related protein
	2080C	Aerobic-type carbon monoxide dehydrogenase, small subunit CoxS/CutS homologs
	1529C	Aerobic-type carbon monoxide dehydrogenase, large subunit CoxL/CutL homologs
12	1271C	Cytochrome bd-type quinol oxidase, subunit 1
	1294C	Cytochrome bd-type quinol oxidase, subunit 2
13	163H	3-polyprenyl-4-hydroxybenzoate decarboxylase
	43H	3-polyprenyl-4-hydroxybenzoate decarboxylase and related decarboxylases
14	1788I	Acyl CoA:acetate/3-ketoacid CoA transferase, alpha subunit
	2057I	Acyl CoA:acetate/3-ketoacid CoA transferase, beta subunit
15	1346M	Putative effector of murein hydrolase
	1380R	Putative effector of murein hydrolase LrgA
16	1653G	ABC-type sugar transport system, periplasmic component
	3839G	ABC-type sugar transport systems, ATPase components
	395G	ABC-type sugar transport system, permease component
	1175G	ABC-type sugar transport systems, permease components
17	1732M	Periplasmic glycine betaine/choline-binding (lipo)protein of an ABC-type transport system
	1125E	ABC-type proline/glycine betaine transport systems, ATPase components
18	1116P	ABC-type nitrate/sulfonate/bicarbonate transport system, ATPase component
	600P	ABC-type nitrate/sulfonate/bicarbonate transport system, permease component
	715P	ABC-type nitrate/sulfonate/bicarbonate transport systems, periplasmic components
19	1203R	Predicted helicases
	1518L	Uncharacterized protein predicted to be involved in DNA repair
20	1638G	TRAP-type C4-dicarboxylate transport system, periplasmic component
	1593G	TRAP-type C4-dicarboxylate transport system, large permease component
21	1640G	4-alpha-glucanotransferase
	58G	Glucan phosphorylase
	296G	1,4-alpha-glucan branching enzyme
	297G	Glycogen synthase
	1523G	Type II secretory pathway, pullulanase PulA and related glycosidases
	448G	ADP-glucose pyrophosphorylase
22	1088M	dTDP-D-glucose 4,6-dehydratase
	1091M	dTDP-4-dehydrorhamnose reductase
	1209M	dTDP-glucose pyrophosphorylase
	1898M	dTDP-4-dehydrorhamnose 3,5-epimerase and related enzymes
23	1129G	ABC-type sugar transport system, ATPase component
	1172G	Ribose/xylose/arabinose/galactoside ABC-type transport systems, permease components
	1879G	ABC-type sugar transport system, periplasmic component
24	2894D	Septum formation inhibitor-activating ATPase
	850D	Septum formation inhibitor
	851D	Septum formation topological specificity factor

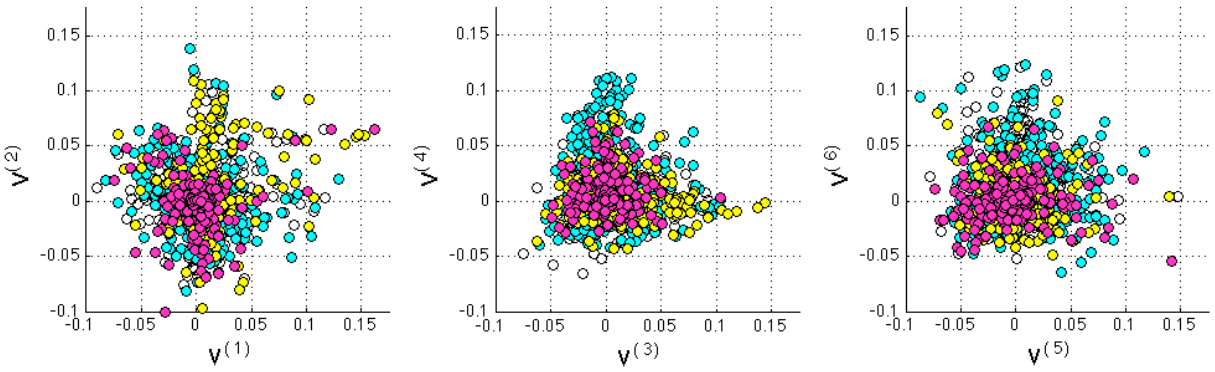

FIG. S 21: Genomic sectors ($k_{\max} = 6$) – COGs are colored as in Fig. S17 according to the functional category to which they belong: cyan for metabolism, yellow for cellular processes, magenta for information processing, and white for poorly characterized genes or genes that belong to multiple categories [12]. The apparent enrichment of yellow COGs along $V^{(2)}$ or cyan COGs along $V^{(4)}$ is quantitatively estimated in Tab. SVII.

Sector	Total number of genes	Info. process.	Cellular process.	Metabolism	p-value for the composition
1	22	1	4	13	0.14
2	56	5	25	17	$5\ 10^{-5}$
3	74	2	35	23	$5\ 10^{-9}$
4	114	8	10	85	$3\ 10^{-9}$
5	50	12	7	24	0.37
6	65	2	4	32	$3\ 10^{-4}$

TAB. S VII: Association between genomic sectors and functional categories – For each sector, defined as the COGs with contribution > 0.05 along one of the 6 components $V^{(k)}$ shown in Fig. S21 and not along any other, we assessed the significance of their content in information processing (magenta), cellular processing (yellow), metabolic (cyan) COGs , which represent respectively around 1/2, 1/4 and 1/4 of the COGs. The p-value is computed from a χ^2-square test with 2 degrees of freedom. 4 of the 6 sectors can be considered to be significantly enriched in COGs of some of these functional categories.