Published online 19 September 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D37–D40***
doi:10.1093/nar/gkn597

DiProDB: a database for dinucleotide properties

Maik Friedel[1], Swetlana Nikolajewa[2], Jürgen Sühnel[1] and Thomas Wilhelm[3,*]

[1]Biocomputing Group, Leibniz Institute for Age Research - Fritz Lipmann Institute, Beutenbergstrasse 11, 07745 Jena, [2]Department of Bioinformatics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany and [3]Theoretical Systems Biology, Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK

Received August 1, 2008; Revised and Accepted September 3, 2008

ABSTRACT

DiProDB (http://diprodb.fli-leibniz.de) is a database of conformational and thermodynamic dinucleotide properties. It includes datasets both for DNA and RNA, as well as for single and double strands. The data have been shown to be important for understanding different aspects of nucleic acid structure and function, and they can also be used for encoding nucleic acid sequences. The database is intended to facilitate further applications of dinucleotide properties. A number of property datasets is highly correlated. Therefore, the database comes with a correlation analysis facility. Authors having determined new sets of dinucleotide property values are invited to submit these data to DiProDB.

INTRODUCTION

Nucleic acid properties are governed by the corresponding nucleotide sequence. More specifically, many properties such as nucleic acid stability, for example, seem to depend primarily on the identity of nearest-neighbour nucleotides (1). The corresponding nearest-neighbour model is also the basis for RNA secondary structure prediction by free-energy minimization (2). It is known that not only thermodynamic but also conformational nucleotide properties may play a role. It has been shown, for example, that promoter locations can be predicted adopting dinucleotide stiffness parameters derived from molecular dynamic simulations (3). Also, curved DNA is known to play a role in prokaryotic gene expression (4). In addition, physical DNA profiles have been used for an improved promoter prediction (5,6). There are numerous other examples. It is, however, beyond the scope of this brief database description to provide a comprehensive overview. Currently, we are developing a Genome Browser that encodes complete eukaryotic or prokaryotic genomes by thermodynamic and conformational dinucleotide properties. In this context, we have collected more than 100 sets of dinucleotide properties from the literature. Currently, there are two related data collections, the PROPERTY DB (srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+LibInfo+-id+1pFZP1TuQpU+-lib+PROPERTY) with about 30 property sets (7) and plot.it (hydra.icgeb.trieste.it/dna/plot_it.html) with about 50 sets (Vlahovicek,K. and Pongor,S., unpublished data). Both of these databases do not include many of the existing datasets and, in addition, it is difficult to trace back the original data sources. Also, both of them are not included in the NAR Database Collection. Therefore, we have set up the database DiProDB, which is aimed to be a one-stop resource for these properties. With DiProDB we want to provide reliable, easily accessible and comprehensive information on dinucleotide properties that may stimulate the application of these data to a diversity of biological problems.

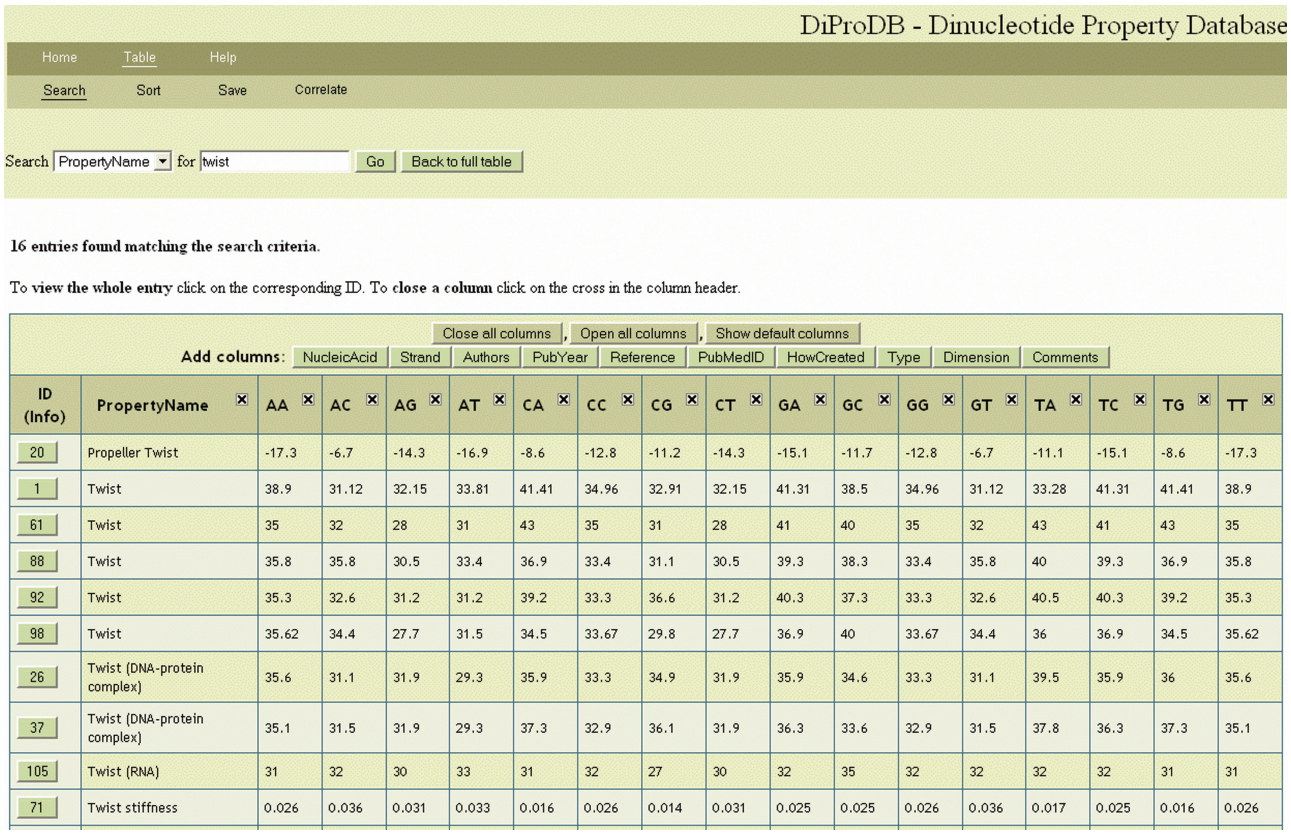
DATABASE CONTENT

DiProDB currently includes 115 dinucleotide datasets. They were collected from the literature and are classified according to nucleic acid type (DNA and RNA), strand information (double or single), how the data were obtained (experimental, theoretical/calculated) and also according to the general type of the dinucleotide property: thermodynamical (e.g. free energy), conformational (e.g. twist) or letter-based (e.g. GC content). We include the letter-based data to demonstrate relations to thermodynamical and conformational properties. Moreover, most of the current motif discovery approaches are letter-based. An example from our work refers to the identification of significant purine–pyrimidine patterns in restriction enzyme binding sites (8). The number of datasets for each category is shown in Table 1. For each dataset, the 16 dinucleotide values, the unit of measurement, the reference, the classification features as well as comments are provided. If a dataset refers to RNA, it is mentioned in the corresponding property name, if the name does not mention a nucleic acid, it always refers to DNA.

*To whom correspondence should be addressed. Tel: +44 1603 255313; Fax: +44 1603 255128; Email: thomas.wilhelm@bbsrc.ac.uk

Table 1. Number of dinucleotide property datasets for each category

Nucleic acid type			Strand information		Mode of property determination		Property type		
DNA	DNA/RNA	RNA	Double	Single	Experimental	Theoretical/calculated	Thermo-dynamical	Conformational	Letter-based
93	7	15	103	12	33	82	34	74	7



ID (Info)	PropertyName	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
20	Propeller Twist	-17.3	-6.7	-14.3	-16.9	-8.6	-12.8	-11.2	-14.3	-15.1	-11.7	-12.8	-6.7	-11.1	-15.1	-8.6	-17.3
1	Twist	38.9	31.12	32.15	33.81	41.41	34.96	32.91	32.15	41.31	38.5	34.96	31.12	33.28	41.31	41.41	38.9
61	Twist	35	32	28	31	43	35	31	28	41	40	35	32	43	41	43	35
88	Twist	35.8	35.8	30.5	33.4	36.9	33.4	31.1	30.5	39.3	38.3	33.4	35.8	40	39.3	36.9	35.8
92	Twist	35.3	32.6	31.2	31.2	39.2	33.3	36.6	31.2	40.3	37.3	33.3	32.6	40.5	40.3	39.2	35.3
98	Twist	35.62	34.4	27.7	31.5	34.5	33.67	29.8	27.7	36.9	40	33.67	34.4	36	36.9	34.5	35.62
26	Twist (DNA-protein complex)	35.6	31.1	31.9	29.3	35.9	33.3	34.9	31.9	35.9	34.6	33.3	31.1	39.5	35.9	36	35.6
37	Twist (DNA-protein complex)	35.1	31.5	31.9	29.3	37.3	32.9	36.1	31.9	36.3	33.6	32.9	31.5	37.8	36.3	37.3	35.1
105	Twist (RNA)	31	32	30	33	31	32	27	30	32	35	32	32	32	32	31	31
71	Twist stiffness	0.026	0.036	0.031	0.033	0.016	0.026	0.014	0.031	0.025	0.025	0.026	0.036	0.017	0.025	0.016	0.026

Figure 1. Screenshot of the DiProDB table displaying search results for the term 'twist' (conformational dinucleotide property) in the property name.

USER INTERFACE

DiProDB displays all data in a single table, see Figure 1. The number and type of columns shown can be customized by the user. When clicking on the ID button in the first column a new page pops up containing all relevant information about the corresponding property. The database entries can be sorted according to three different criteria. There is also a search option for all or for specific columns. The complete table or parts of it can be saved as text file or in a format directly importable into the Genome Browser mentioned in the Introduction section. The DiProDB website contains a Submit button, where users can submit new property datasets.

DATA ANALYSES

The DiProDB website contains a Correlate option, where users can calculate Pearson's or Spearman's rank correlation coefficients for all or selected properties. This allows easy identification of dependencies between different dinucleotide properties. As an example in Figure 2, Spearman's correlation data are shown for five different datasets quantifying the twist in B-DNA. All datasets are clearly correlated to each other. However, the extent of correlation is rather different. Correlation coefficients >0.58 are considered as statistically significant ($P < 0.01$, t-test).

Based on these correlations, we have done different hierarchical clustering analyses to get a deeper insight into the overall correlation of the datasets. Figure 3 shows a single linkage hierarchical clustering of all 23 B-DNA double-strand thermodynamical properties together with the three-dinucleotide letter-based quantities GC content, purine (GA) content and keto (GT) content. This clustering is based on the distance measure $1-|r_{\mathrm{Pearson}}|$, because it is just the absolute value of the correlation, which indicates whether two properties contain similar information.

Save table	1 Twist	61 Twist	88 Twist	92 Twist	98 Twist
1 Twist		0.7630	0.5959	0.7261	0.5974
61 Twist	0.7630		0.8633	0.9085	0.7597
88 Twist	0.5959	0.8633		0.7924	0.8958
92 Twist	0.7261	0.9085	0.7924		0.6344
98 Twist	0.5974	0.7597	0.8958	0.6344	

Figure 2. Pearson's correlation coefficients for five sets of twist angles. ID (Ref.): 1 (9), 61 (10), 88 (11), 92 (12) and 98 (13). Correlation coefficients >0.8 are coloured in green.

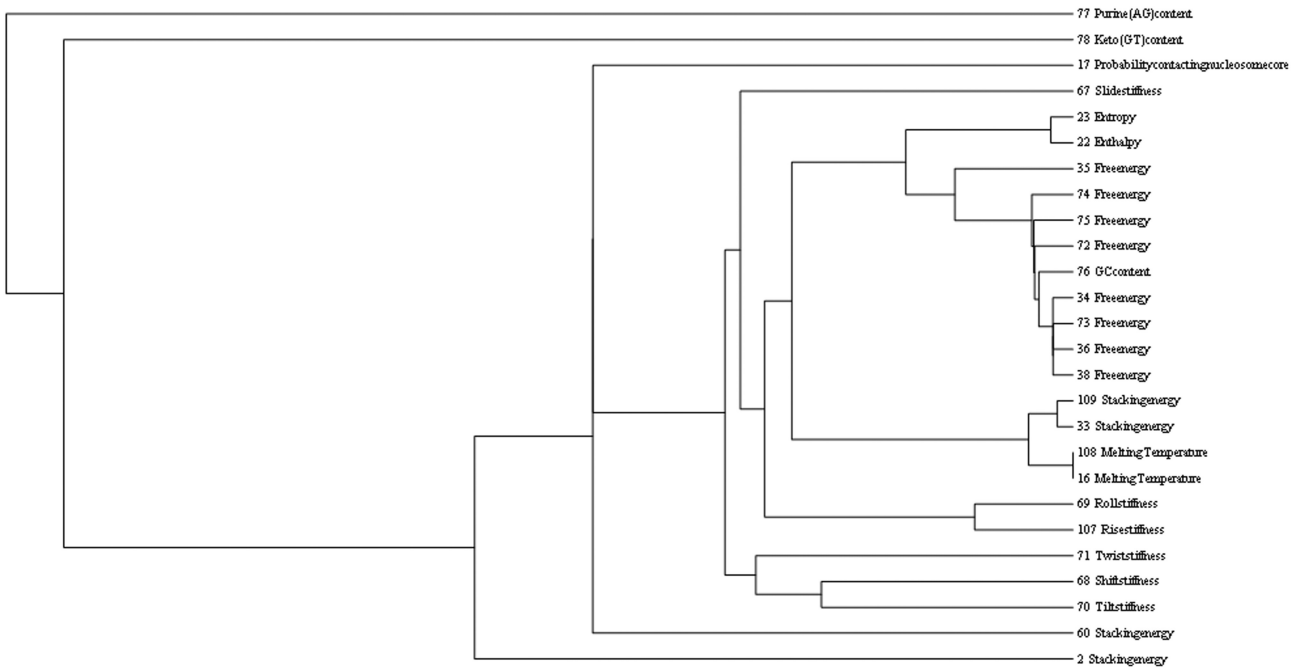


Figure 3. Hierarchical clustering of all 23 B-DNA double-strand physicochemical properties and the three-dinucleotide letter-based quantities GC content, purine (GA) content and keto (GT) content. The property sets are designated by their IDs and names.

Other correlation measures like Spearman or Kendall-Tau give very similar results. It can be seen that all free-energy data contain more or less the same information and that this is basically equivalent to the GC content. This is very likely due to the simple fact that GC pairs have three H-bonds instead of two in AT base pairs. The complete single-linkage hierarchical clustering of all 115 properties is given in the Supplementary Material (Table 2), where also a corresponding Ward clustering (14) is shown. The latter one shows a separation between a free energy/entropy/enthalpy/stacking energy/melting temperature cluster and another cluster containing all the conformational datasets. The complete single linkage clustering reveals that the most uncorrelated dinucleotide properties are direction, inclination, twist–rise (conformational), stacking energy, tilt, shift, propeller twist and rise.

In order to gain more insights into the data, we performed two principal component analyses (PCA) (15). The complete data of 115 properties for 16 dinucleotides corresponds to 115 points in 16-dimensional space (or 16 points in 115-dimensional space). PCA helps to reveal the internal structure of such high-dimensional data by providing lower dimensional pictures of the 'cloud' in coordinates corresponding to maximum variance of the data (http://en.wikipedia.org/wiki/Principal_components_analysis). The cloud of all 115 properties in the first two principal components (PCs, the new coordinates) is shown in Figure 4. Only the most uncorrelated property 'direction' lies outside the shown region: $(PC1,PC2)_{Direction} = (0.1,1.6)$ (the complete figure containing direction and a PC1–PC3 projection are given in the Supplementary Material; note also that only the first three PCs carry relevant information: PC1 78.5%, PC2 16.9%, PC3 3.3%). The other two outliers are melting temperature and persistence length. This indicates that especially these three properties carry information quite different from the others. Note that the latter two properties are not amongst the outliers according to the above mentioned single linkage clustering, because each one has (at least) one better correlation to other datasets (melting temperature to stacking energy, and persistence length to tilt–shift). Figure 4 also indicates three clusters containing all other properties, one stacking energy/entropy cluster, a twist cluster and the central main cluster.

Table 2. Content of supplementary material

Figure S1	Single linkage hierarchical clustering of 115 dinucleotide properties.
Figure S2	Ward hierarchical clustering of 115 dinucleotide properties.
Figure S3	115 dinucleotide properties in the first two principal components.
Figure S4	115 dinucleotide properties in the first and third principal component.
Figure S5	The 16 dinucleotides in the first two principal components.
Table S1	Percentage of importance of the 15 PCs carrying $>10^{-14}$% of information.
Table S2	Percentage of importance of the first 10 dinucleotide properties in the first 15 PCs in decreasing order.
Table S3	Involvement of the 10 most important dinucleotide properties in the PCs 1–15.

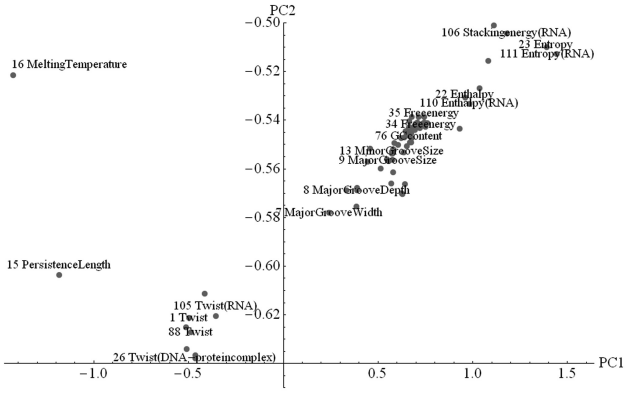


Figure 4. All dinucleotide properties plotted in the first two PCs. A few of them are designated by property name and ID.

Finally, we also performed a PCA calculating the 115 principal components for the 16 dinucleotides. The first 15 PCs carry information (23%, 21%, 14%, 12%, 6%, etc.), roughly indicating that about this number of low correlated properties is needed to represent all information of the complete set of 115 properties. The Supplementary Material also contains a corresponding PC1–PC2 plot, together with all detailed information about the performed PCAs.

OUTLOOK

So far the DiProDB database contains 115 sets of dinucleotide properties. In the future, this number is to be increased. We also invite other authors to submit their measured or calculated dinucleotide properties to DiProDB.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Friedrich Haubensak for setting up the database and to Rolf Hühne for helpful comments on the database layout.

FUNDING

Funding for open access charge: Biotechnology and Biological Sciences Research Council (BBSRC)IFR Core Strategic Grant.

Conflict of interest statement. None declared.

REFERENCES

1. SantaLucia,J. Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbour thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
2. Mathews,D.H. and Turner,D.H. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, **16**, 270–278.
3. Goñi,J.R., Pérez,A., Torrents,D. and Orozco,M. (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.*, **8**, R263.
4. Pérez-Martín,J., Rojo,F. and de Lorenzo,V. (1994) Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. *Microbiol. Rev.*, **58**, 268–290.
5. Abeel,T., Saeys,Y., Rouzé,P. and Van de Peer,Y. (2008) ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinformatics*, **24**, i24–i31.
6. Florquin,K., Saeys,Y., Degroeve,S., Rouzé,P. and Van de Peer,Y. (2005) Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Res.*, **33**, 4255–4264.
7. Ponomarenko,J.V., Ponomarenko,M.P., Frolov,A.S., Vorobyev,D.G., Overton,G.C. and Kolchanov,N.A. (1999) Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics*, **15**, 654–668.
8. Nikolajewa,S., Beyer,A., Friedel,M., Hollunder,J. and Wilhelm,T. (2005) Common patterns in type II restriction enzyme binding sites. *Nucleic Acids Res.*, **33**, 2726–2733.
9. Karas,H., Knüppel,R., Schulz,W., Sklenar,H. and Wingender,E. (1996) Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *Comput. Appl. Biosci.*, **12**, 441–446.
10. Pérez,A., Noy,A., Lankas,F., Luque,F.J. and Orozco,M. (2004) The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res.*, **32**, 6144–6151.
11. Gorin,A.A., Zhurkin,V.B. and Olson,W.K. (1995) B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.*, **247**, 34–48.
12. Suzuki,M., Yagi,N. and Finch,J.T. (1996) Role of base-backbone and base-base interactions in alternating DNA conformations. *FEBS Lett.*, **379**, 148–152.
13. Shpigelman,E.S., Trifonov,E.N. and Bolshoy,A. (1993) CURVATURE: software for the analysis of curved DNA. *Comput. Appl. Biosci.*, **9**, 435–440.
14. Ward,J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.
15. Pearson,K. (1901) On lines and planes of closest fit to systems of points in space. *Philos. Magazine*, **2**, 559–572.