Available online at www.elixirpublishers.com (Elixir International Journal)

Discrete Mathematics

Elixir Dis. Math. 53A (2012) 12208-12211

ISSN: 2229-712X

Mathematical representation of high-resolution protein structures using graph theory

M.El-Ghoul[1], A.El-Guoshy[2], F.El-Fiki[2] and A.El-Refy[2]

[1]Department of Mathematics, Faculty of Science, Tanta University.

[2]Biotechnology Department, Faculty of Agriculture, Al-Azhar University.

ARTICLE INFO

Article history:

Received: 24 November 2011;

Received in revised form:

22 November 2012;

Accepted: 10 December 2012;

Keywords

Protein,

Amino acid,

Graph theory,

Matrices.

ABSTRACT

In this paper we developed new matrices (angle matrix ,connection matrix and connection angle matrix) in addition to a set of matrices (adjacent matrix, incidence matrix) that previously used to introduce the mathematical representation of high-resolution protein structures an accurate representatin that facilitate analysis of its structures.

Introduction

Definition and background:-

- **Abstract graphs:** An abstract graphs G is a diagram consisting of a finite non empty set of the elements, called "vertices" denoted by V (G) together with a set of unordered pairs of these elements, called "edges" denoted by E (G). The set of vertices of the graph G is called "the vertex set of G" and the list of edges is called "the edge – list of G" [Gibbons, 1995; Giblin, 1977].

- **Adjacency and incidence:** let v and w be vertices of a graph. If v and w are joined by an edge e. then v and w are said to be adjacent. Moreover, v and w are said to be incident with e, and e is said to be incident with v and w [Wilson, 1972].

- **The adjacency matrix:** let G be a graph without loops, with n-vertices labeled 1, 2, 3, ..., n. The adjacency matrix A(G) is the nxn matrix in which the entry in row i and column j is the length of edge in angstroms if the vertices i and j are joining and 0 otherwise. [Modification: Wilson, 1972].

- **The incidence matrix:** let G be a graph without loops, with n-vertices labeled 1, 2, 3, . . . n and m edges labeled 1, 2, 3, ..., m. the incidence matrix I (G) is the n x m matrix in which the entry in row i and column j is l if vertex i is incident with edge j and 0 otherwise [Gross, 1987; Wilson, 1990].

- **Angle matrix:** let G be a graph without loops, with n-vertices labeled 1, 2, 3, . . . n and m edges labeled 1, 2, 3, ..., m. the angle matrix g(G) is the m x m matrix in which the entry in row i and column j is the angle in degrees if edge i and j are incident with the same vertex and 0 otherwise.(New matrix)

- **Connection matrix:** let g1 be a directed subgraph without loops, with n1-vertices labeled 1, 2, 3. . . n1 and g2 be another directed subgraph without loops, with n2-vertices labeled 1, 2, 3, ..., n2. The connection matrix C (g1, g2) is the n1xn2 matrix in which the entry in row i and column j is the length of edge in angstroms if the vertices i and j are joining and 0 otherwise. [New matrix].

- **Connection Angle matrix:** let g1 be a directed weighted subgraph without loops, with n1-vertices labeled 1, 2, 3. . . . n1, m1- edges labeled 1, 2, 3…m1, g2 be another directed weighted subgraph without loops, with n2-vertices labeled 1, 2, 3. . . n2, m2- edges labeled 1, 2, 3…m2 and E is a directed connector between g1, g2. the connection angle matrix CL(G) is the m1 x E matrix or E x m2 in which the entry in row i and column j is the angle in degrees if edge i and j are incident with the same vertex and 0 otherwise.(New matrix)

Main results

Now we will represent the Mean values of Main-chain bond lengths and bond angles extracted from PROCHECK Operating Manual Appendix A Stereochemical parameters TABLE A.1 (http://www.ebi.ac.uk/thornton srv/software/PROCHECK/manual/manappa.html) as observed in small molecules (Engh & Huber, 1991) that derived from high-resolution protein structures using graph matrices (Figure 1).

Definition

Representation of high-resolution protein structure backbone using a directed weighted graph (G) in consideration atoms as vertices and chemical bonds between atoms as edges in 2 steps.

The First step:-

Representation of each amino acid residue backbone (except glycine and proline) as undirected weighted sub graph (g) all in protein primary structure as a directed weighted graph (G) using 2 matrices.

A $(g)_{all}$ =

N Cα C O	N	Cα	C	O
	0	1.458	0	0
	1.458	0	1.525	0
	0	1.525	0	1.231
	0	0	1.231	0

Where N, Cα, C, O represent the vertices of subgraph $(g)_{all}$ and values in cells represent the length of edges in angstrom.

$I\ (g)_{all}$ =

Tele:

E-mail addresses: hendelmorsy@yahoo.com

N Cα C O	N Cα	CαC 0	CO
	1	1 1 0	0
	1		0
	0		1
	0		1

Where N, Cα, C,O represent the vertices and N Cα, CαC, CO represent the edges of subgraph (g) $_{all}$ and values in cells represent the existence or absence of connection.

L (g) $_{all}$ =

	N Cα	CαC	CO
N Cα	0	111.2	0
CαC	111.2	0	120.8
CO	0	120.8	0

Where N Cα, CαC, CO represent the edges of aa backbone graph (all aa except glycine and proline) and values in cells represent the angle values in degrees between two edges.

Representation of the amino acid residue glycine backbone as subgraph (g)$_g$

A (g) $_g$ =

N Cα C O	N	Cα	C	O
	0	1.451	0	0
	1.451	0	1.516	0
	0	1.516	0	1.231
	0	0	1.231	0

Where N, Cα, C, O represent the vertices of subgraph (g)$_g$ and values in cells represent the length of edges in angstrom

I (g) $_g$ =

N Cα C O	N Cα	CαC 0	CO
	1	1 1 0	0
	1		0
	0		1
	0		1

Where N, Cα, C,O represent the vertices and N Cα, CαC, CO represent the edges of subgraph (g)$_g$ and values in cells represent the existence or absence of connection.

L (g) $_g$=

	N Cα	CαC	CO
N Cα	0	112.5	0
CαC	112.5	0	120.8
CO	0	120.8	0

Where N Cα, CαC, CO represent the edges of glycine backbone graph and values in cells represent the angle values in degrees between two edges.

Representation of the amino acid residue proline backbone as subgraph (g)$_p$.

A (g) $_p$=

N Cα C O	N	Cα	C	O
	0	1.466	0	0
	1.466	0	1.525	0
	0	1.525	0	1.231
	0	0	1.231	0

Where N, Cα, C, O represent the vertices of subgraph (g) $_p$ and values in cells represent the length of edges in angstrom

I (g) $_p$ =

N Cα C O	N Cα	CαC 0	CO
		1	0 0
	1	1 0	1 1
	1		
	0		
	0		

Where N, Cα, C,O represent the vertices and NCα, CαC, CO represent the edges of subgraph (g) $_p$ and values in cells represent the existence or absence of connection.

L (g) $_p$=

	N Cα	CαC	CO
N Cα	0	111.8	0
CαC	111.8	0	120.8
CO	0	120.8	0

Where N Cα, CαC, CO represent the edges of proline backbone graph and values in cells represent the angle values in degrees between two *edges.*

The second step: -

Representation of each connection among (g) $_{all}$, (g) $_g$ and (g) $_p$ subgraphs using a connection matrix and angle matrix.

C (g_{any},g_{all}) =

N Cα C O	N	Cα	C	O
	0	0	0	0
	0	0	0	0
	1.329	0	0	0
	0	0	0	0

Figure 1
Main-chain bond lengths and bond angles extracted from PROCHECK Operating Manual Appendix A Stereochemical parameters TABLE A.1

a. Bond lengths

Bond	X-PLOR labelling		Value	sigma
C-N	C-NH1	(except Pro)	1.329	0.014
	C-N	(Pro)	1.341	0.016
C-O	C-O		1.231	0.020
Calpha-C	CH1E-C	(except Gly)	1.525	0.021
	CH2G*-C	(Gly)	1.516	0.018
Calpha-Cbeta	CH3E-CH1E	(Ala)	1.521	0.033
	CH1E-CH1E	(Ile,Thr,Val)	1.540	0.027
	CH1E-CH2E	(the rest)	1.530	0.020
N-Calpha	NH1-CH1E	(except Gly,Pro)	1.468	0.019
	NH1-CH2G*	(Gly)	1.451	0.016
	N-CH1E	(Pro)	1.466	0.015

b. Bond Angles

Angle	X-PLOR labelling		Value	sigma
C-N-Calpha	C-NH1-CH1E	(except Gly,Pro)	121.7	1.8
	C-NH1-CH2G*	(Gly)	120.6	1.7
	C-N-CH1E	(Pro)	122.6	5.0
Calpha-C-N	CH1E-C-NH1	(except Gly,Pro)	116.2	2.0
	CH2G*-C-NH1	(Gly)	116.4	2.1
	CH1E-C-N	(Pro)	116.9	1.5
Calpha-C-O	CH1E-C-O	(except Gly)	120.8	1.7
	CH2G*-C-O	(Gly)	120.8	2.1
Cbeta-Calpha-C	CH3E-CH1E-C	(Ala)	110.5	1.5
	CH1E-CH1E-C	(Ile,Thr,Val)	109.1	2.2
	CH2E-CH1E-C	(the rest)	110.1	1.9
N-Calpha-C	NH1-CH1E-C	(except Gly,Pro)	111.2	2.8
	NH1-CH2E*-C	(Gly)	112.5	2.9
	N-CH1E-C	(Pro)	111.8	2.5
N-Calpha-Cbeta	NH1-CH1E-CH3E	(Ala)	110.4	1.5
	NH1-CH1E-CH1E	(Ile,Thr,Val)	111.5	1.7
	N-CH1E-CH2E	(Pro)	103.0	1.1
	NH1-CH1E-CH2E	(the rest)	110.5	1.7
O-C-N	O-C-NH1	(except Pro)	123.0	1.6
	O-C-N	(Pro)	122.0	1.4

Where the vertical axe (g1) represents (g) $_{all}$ or (g) $_g$ or (g) $_p$ subgraph vertices and the horizontal one (g2) represents (g) $_{all}$

subgraph vertices. And values in cells represent the length of edges in angstrom.

Figure 2
All amino acid residue backbone (except glycine and proline)

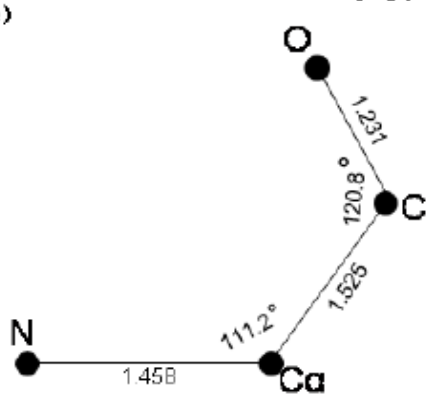


Figure 3
glycine amino acid residue backbone .

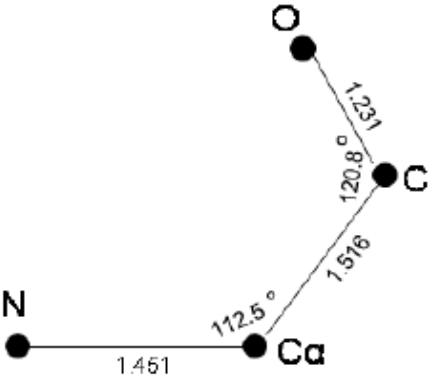
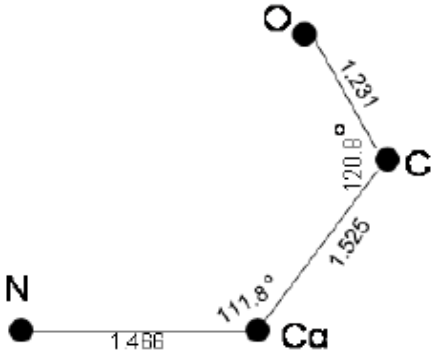


Figure 4
proline amino acid residue backbone



CL (g_{any}, CN & CN, g_{all}) =

g any	CN		g all
NCα	0	121.7	NCα
CαC	116.2	0	CαC
OC	123.0	0	OC

Where N Cα, CαC, OC represents the edges of g1, g2 directed subgraph, g2 represents $(g)_{all}$,CN represent the directed connection between g1,g2 and values in cells represent the angle values in degrees between two edge.

	N	Cα	C	O
	0	0	0	0
	0	0	0	0
N Cα	1.341	0	0	0
C O	0	0	0	0

Where the vertical axe (g1) represents $(g)_{all}$ or $(g)_{g}$ or $(g)_{p}$ subgraph vertices and the horizontal one (g2) represents $(g)_{g}$ subgraph vertices. And values in cells represent the length of edges in angstrom.

CL (g_{any}, CN & CN, g_{g}) =

g any	CN		g_{g}
NCα	0	120.6	NCα
CαC	116.2	0	CαC
OC	123.0	0	OC

Where N Cα, CαC, OC represent the edges of g1, g2 directed subgraph, g2 represents $(g)_{g}$,CN represent the directed connection between g1, g2 and values in cells represent the angle values in degrees between two edges.

C (g_{any},g_{p}) =

N Cα C O	N	Cα	C	O
	0	0	0	0
	0	0	0	0
	1.341	0	0	0
	0	0	0	0

Where the vertical axe (g1) represents $(g)_{all}$ or $(g)_{g}$ or $(g)_{p}$ subgraph vertices and the horizontal one (g2) represents $(g)_{p}$ subgraph vertices. And values in cells represent the length of edges in angstrom

CL(g_{any}, CN & CN, g_{g}) =

g any	CN		g_{g}
NCα	0	122.6	NCα
CαC	116.9	0	CαC
OC	122.0	0	OC

Where N Cα, CαC, O C represent the edges of g1,g2 directed subgraph , g2 represents $(g)_{p}$,CN represent the directed connection between g1,g2 and values in cells represent the angle values in degrees between two edges.

Figure 5
Any amino acid residue backbone and All amino acid residue backbone (except glycine and proline) dipeptide.

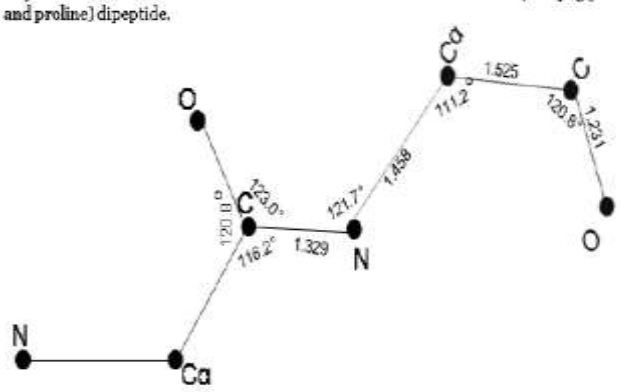


Figure 6
Any amino acid residue backbone and glycine amino acid residue backbone dipeptide.

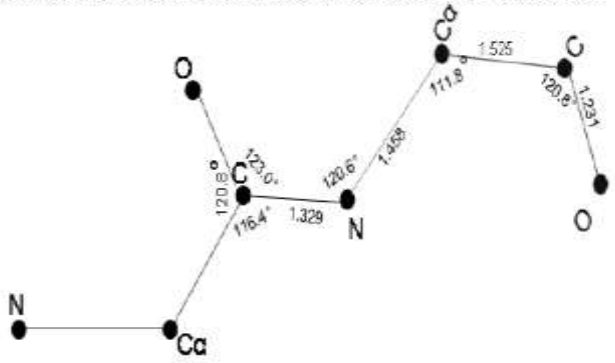
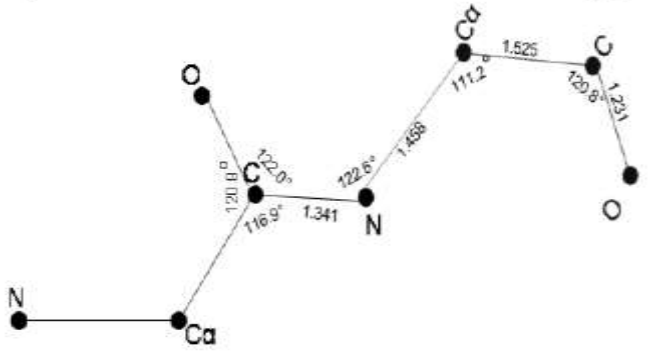


Figure 7
Any amino acid residue backbone and Proline amino acid residue backbone dipeptide.



References

1. Gibbons, A. (1995). Algorithmic graph theory. Cambridge University Press, Cambridge, UK.
2. Giblin, P.J. (1977). Graphs, surfaces and homology, an introduction to algebraic topology. Chapman and Hall. Ltd, London 1977.
3. Gross, J.L. and Tucker, T.W. (1987). Topological graph theory. Jon Wiley & Sons, Inc, Canada 1987.
4. Wilson, R.J. (1972). Introduction to graph theory. Olivar& Boyed, Edinburgh.
5. Wilson, R.J. and Watkins, J.J. (1990). Graphs, an introductory approach, a first course in discrete mathematics. Jon Wiley & Sons Inc, Canada.