Building Geometrical Models for Biological Molecules

by

Haruna Katayama

Bachelor of Science
University of South Carolina 2000

Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in the

Department of Mathematics

University of South Carolina

2003

Department of Mathematics Director of Thesis	Department of Mathematics Second Reader

Dean of The Graduate School

Dedication

To my Father, Who is in heaven.

To my Brother, through Whom all things came into being.

To my Helper, Who teaches me all things.

I will lift up my eyes to the mountains;
From where shall my help come?
My help comes from the LORD,
Who made heaven and earth.
Psalm 121:1,2

Acknowledgements

First, I would like to thank Dr. Daniel Dix, my thesis advisor, who has patiently guided me through this long process of completing the thesis, encouraging me through the Scripture when I felt that I could not go on any more and bearing with me in my ignorance and frustration. I recognize and thank his family for the sacrifices that they made for him to do so.

Also, I am grateful to Dr. László Székely, my second reader, for being a second reader, waiting quietly without pressuring at all till I was able to hand the thesis for his reading, and his insightful suggestion and uplifting words.

I would like to express special thanks to Dr. George McNulty, who has been not only the technical support, but more than that, has been a support throughout my time here in this school, even from my undergraduate time, and encouraged me to go on to graduate school.

I am thankful for Dr. Anton Schep, the director of the Department of Mathematics, for his advise and guidance and help in many ways, going beyond his duty as to look for an office space for my last semester.

To all the staff in the Department, I give thanks for what they have done and are doing and the care that they show to us. And many thanks to my fellow graduate students and office mates for various ways of support and kindness.

I am deeply grateful for my family and the Family for their deep love, continous support, care and labor, taking part in this project for His glory. By His power through their labor, this servant was enabled to persevere. His grace was and is sufficient. May He reward and bless each one.

And most of all, I give thanks to my Lord, my Creator, the living God for sustaining and teaching me and for His great love.

Acknowledgements for the softwares used in this project:

- Convert was written by Jason Rogers.
- IMIMOL was written by Scott Johnson with support from the Industrial Mathematics Institute at the University of South Carolina.
- MolProbity was written by Ian Davis with support from the Richardson Lab at Duke University.
- RasMol was written by Roger Sayle with support from Glaxo Wellcome Research and Development.
- VMD was written by the Theoretical Biophysics group and the Beckman Institute at the University of Illinois, Urbana, IL, with support from the NSF and NIH.
- VMD-IMI was written by Matthew Heilsberg and Matt Elder with support from the Industrial Mathematics Institute at the University of South Carolina.

Contents

Dedication ii

Acknowledgements iii

List of Figures vii

List of Tables x

Introduction 1

Chapter 1 GZ-systems 3

1.1. Configuration and Conformation 3

1.2. Definitions and Notations 6

1.3. The Definition of Generalized Z-systems 15

1.4. Conformations and Labeled Z-systems 18

1.5. Coordinate Transformation Matrix 24

1.6. Tethering and Gluing 27

1.7. Conformations and Labeled GZ-systems 32

1.8. Example of a GZ-System 36

Chapter 2 Light-Harvesting Complex 46

2.1. Biological Setting 46

2.2. Chemical Nature of Building Blocks 49

Chapter 3 Homology Modeling 58

3.1. Homology Modeling Overview 58

3.2. Goal and Principles 59

3.3. LH-II from *Rb. sphaeroides* 61

Chapter 4 Building a Model 64

4.1. The Z-system of A Protein Chain 64

4.2. Z-systems of BCLs, Carotenoids, and Lipids 71

4.3. A Z-system for a PC 72

4.4. A Labeled Z-system 76

4.5. PC for *sph.* 81

4.6. Conformation Assignment 83

4.7. A Z-System of LH-II 88

4.8. Complex Conformational Adjustment 90

Chapter 5 Evaluation 92

5.1. Summary 92

5.2. Discussion 94

5.3. Conclusion 97

Appendix A: The Structure of LH-II 98

Appendix B: The Z-systems 103

Glossary 119

Bibliography 123

List of Figures

1 Examples of Graph 7

2 Example of a GZ-system 17

3 Positive Orientation of $\mathbf{u}, \mathbf{v}, \mathbf{w}$ 20

4 Negative Orientation of $\mathbf{u}, \mathbf{v}, \mathbf{w}$ 20

5 Two Orientations 21

6 Tethering a 2-Site to a 2-Site 27

7 Tethering a 1-Site to a 2-Site 28

8 Component/Linkage Tree for the GZ-system 37

9 Z-system of Na and H_2O 38

10 Example of Site Graph 39

11 Sequences of Sites and T-Matrices 40

12 Tethered Z-system of OH and H_2O 42

13 Z-system of Carbonic Acid and H_2O 44

14 Hierarchy Tree of LH-II 48

15 Hierarchy Tree of PC 48

16 Phosphatidylethanolamine 50

17 Examples of Fatty Acids 50

18 Glycine Amino Acid (Backbone with an H side chain) 51

19 A Chain of Two Glycine Amino Acids 51

20 Hydrophobic Amino Acids 53

21 Polar Amino Acids 54

22 Charged Amino Acids 54

23 Hierarchy Tree for Protein 55

24 Chemical Structure of BCL a 55

25 Chemical Structures of Carotenoids 57

26 The Principles Governing the Experiment 59

27 The Sequence Alignment 62

28 Z-System of "Backbone" 65

29 Z-System of the Leucine Side Chain 66

30 Z-system of Glued Backbone 68

31 Z-System of Leucine Side Chain Glued to Backbone 69

32 Selection of Sites 73

33 Carboxyl on N Terminus 74

34 Tethering of the Reference Atoms and PCs 88

35 Nine PCs 98

36 The Structure of a PC 99

37 The Photosynthetic System in Cell Membrane 100

38 The Ring of B850 BCLs 100

39 B800 Plane 101

40 A Heterodimer and SPOs 102

41 The Position of SPO2 102

42 Backbone, Alanine, and Arginine 104

43 Asparagine, Aspartic Acid, Cysteine, and Glutamic Acid 105

44 Glutamine and Histidine 106

45 Isoleucine and Leucine 107

46 Lysine, Methionine, and Phenylalanine 108

47 Proline, Serine, Threonine, and Valine 109

48 Tryptophan and Tyrosine 110

49 BCL 1 111

50 BCL 2 112

51 BCL 3 and BCL 4 113

52 SPO 1 and RG 1 114

53 SPO 2 and RG 2 115

54 SPO 3 and RG 3 116

55 PEL 1 117

56 PEL 2 118

List of Tables

1 Example of an Incidence Table 7

2 Amino Acid Names and Abbreviations 52

3 Rotamer Table for Leucine [19] 67

4 Non-Rotameric Conserved Residues 84

5 Nonconserved Residues 85

6 Summary of Rotamerizability 92

7 Inter-Pigment Hydrogen Bonds 94

Introduction

The light-harvesting complex (LHC), a part of the photosynthetic system in purple bacteria has been studied for its simplicity of structure in relation to its known function and as a means to understand transmembrane multiprotein complexes in general.

Despite of the fact that they have different sequences of amino acids which make up the proteins, LHCs in different species have remarkably similar structures [14]. In the study of proteins, it gives a great deal of understanding into the function of the protein to know the structure of the protein. Therefore when the structure is not known for a particular protein under study, the structure of another protein is sometimes used to build a model for that protein. In such case, we assume that the more similar the functions and the amino acid sequences are, the more similar the structures would be. (A method of model building based on this assumption is called homology modeling.) So it is natural to attempt build LHC in one species from the structure of LHC in another species, especiallly if the sequences are similar. To complete the model building process, the model is simulated to relax the structure, that is, to lower the chemical energy caused by unnatural shapes of amino acids, so that it will be closer to the structure in a natural environment. Then the questions arise: is it possible to change the amino acids and other necessary parts from one species to another and still keep the exact structure for the other major parts? How much structural change is allowed to carry over the exact major structure of one species to another?

To explore such questions, with the hope of this being the first step toward answering them, we have conducted an experiment of building a model of LH-II, one type of LHC, in *Rhodobactor (Rb.) sphaeroides* using the LH-II structures of *Rhodopseudomonas (Rps.) acidophila* and *Rhodospirillum (Rs.) molischianum.*

In Chapter 1, the mathematical background is explained for a method of model building using internal coordinates based on Z-systems, along with some concrete examples.

In Chapter 2, we introduce the environment surrounding LH-II and the structure of LH-II as well as some chemistry involved in building the model of LH-II.

Chapter 3 presents the method of homology modeling briefly and describes how it is applied in our experiment. Also, the goal and principles that we followed in the experiment are stated in this chapter.

In the following chapter, Chapter 4, we explain how we have constructed the Z-system to build the model of LH-II from *Rb. sphaeroides.*

Finally, Chapter 5 is the summary and the evaluation of our experiment including what we have learned and improvements that should be made to facilitate more complex model building projects in the future.

Chapter 1

GZ-systems

1.1. Configuration and Conformation

Any movement of a rigid object in three dimensional space can be expressed as a translation, a rotation, or a combination of both. So let $\mathbb{R}^3$ be a group of translations and $\mathrm{SO}(3) = \{A \in \mathbb{R}^{3\times 3} \mid A^T A = I \text{ and } \det A = 1\}$ be the set of rotation matrices where A^T denotes the transpose of a matrix A. It is easy to check that SO(3) is a group. Also note that for $A \in \mathrm{SO}(3)$, $A^T A = AA^T = I$, and that SO(3) does not include a reflection matrix since for a reflection matrix $A = I - 2\mathbf{u}\mathbf{u}^T$ where $\|\mathbf{u}\| = 1$, $\det A = -1$. Let $G_a = \mathbb{R}^3 \times \mathrm{SO}(3)$, called the group of three-dimensional **rigid motions**, be equipped with the binary operation $(\mathbf{b}_1, A_1) \cdot (\mathbf{b}_2, A_2) = (\mathbf{b}_1 + A_1\mathbf{b}_2, A_1A_2)$ for all $(\mathbf{b}_1, A_1), (\mathbf{b}_2, A_2) \in G_a$. The identity of G_a is $(\boldsymbol{\theta}, I)$ where $\boldsymbol{\theta} \in \mathbb{R}^3$ is a zero column vector, and the inverse of $(\mathbf{b}, A)$ is $(-A^T\mathbf{b}, A^T)$. Define a mapping $G_a \times \mathbb{R}^3 \to \mathbb{R}^3 : (g, \mathbf{x}) \mapsto g \cdot \mathbf{x}$ where $(\mathbf{b}, A) \cdot \mathbf{x} = \mathbf{b} + A\mathbf{x}$ for $(\mathbf{b}, A) \in G_a$ and $\mathbf{x} \in \mathbb{R}^3$. This mapping defines a left-action of G_a on $\mathbb{R}^3$ since we have $[(\mathbf{b}_1, A_1)(\mathbf{b}_2, A_2)] \cdot \mathbf{x} = (\mathbf{b}_1, A_1)[(\mathbf{b}_2, A_2) \cdot \mathbf{x}]$ for any $(\mathbf{b}_1, A_1), (\mathbf{b}_2, A_2) \in G_a$ and $\mathbf{x} \in \mathbb{R}^3$. G_a also acts on $\mathbb{R}^{3\times 4}$ from the left with the rule $(\mathbf{b}, A) \cdot (\mathbf{x}, X) = (\mathbf{b} + A\mathbf{x}, AX)$ where $X \in \mathbb{R}^{3\times 3}$.

Now define G_p to be a group of all 4×4 matrices of the form $\left(\begin{smallmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b} & A \end{smallmatrix}\right)$, where $(\mathbf{b}, A) \in G_a$ and $\boldsymbol{\theta} \in \mathbb{R}^3$ is a zero column vector. The group operation of G_p is ordinary matrix multiplication, and G_p acts on $\mathbb{R}^{3\times 4}$ from the right via ordinary matrix multiplication. The right action of G_p commutes with the left action of G_a. So we would get the same result whether we first let $g \in G_a$ act on a matrix $C \in \mathbb{R}^{3\times 4}$

from the left, and then let $M \in G_p$ act on the result from the right, or the other way around, that is, $(g \cdot C)M = g \cdot (CM)$. To see this, let $(\mathbf{b}, A) \in G_a$, $\left(\begin{smallmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b}' & A' \end{smallmatrix}\right) \in G_p$, and $(\mathbf{x}, X) \in \mathbb{R}^{3\times 4}$ where $\mathbf{x} \in \mathbb{R}^3$ and $X \in \mathbb{R}^{3\times 3}$. Then:

$$\begin{aligned}
[(\mathbf{b}, A) \cdot (\mathbf{x}, X)] \begin{pmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b}' & A' \end{pmatrix} &= (\mathbf{b} + A\mathbf{x}, AX) \begin{pmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b}' & A' \end{pmatrix} \\
&= (\mathbf{b} + A\mathbf{x} + AX\mathbf{b}', AXA')
\end{aligned}$$

and

$$\begin{aligned}
(\mathbf{b}, A) \cdot \left[(\mathbf{x}, X) \begin{pmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b}' & A' \end{pmatrix} \right] &= (\mathbf{b}, A) \cdot (\mathbf{x} + X\mathbf{b}', XA') \\
&= (\mathbf{b} + A(\mathbf{x} + X\mathbf{b}'), AXA') \\
&= (\mathbf{b} + A\mathbf{x} + AX\mathbf{b}', AXA');
\end{aligned}$$

thus,

$$[(\mathbf{b}, A) \cdot (\mathbf{x}, X)] \begin{pmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b}' & A' \end{pmatrix} = (\mathbf{b}, A) \cdot \left[(\mathbf{x}, X) \begin{pmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b}' & A' \end{pmatrix} \right].$$

Denote the set of all mapping from a set S to a set T by T^S, and let $\mathcal{N}$ be a set of atom names of all the atoms in a biological system under study. Then a mapping $R \in (\mathbb{R}^3)^{\mathcal{N}}$, $R : i \mapsto \mathbf{R}_i$ from $\mathcal{N}$ to $\mathbb{R}^3$, is called a **configuration** of the system. For $(\mathbf{b}, A) \in G_a$ and $R \in (\mathbb{R}^3)^{\mathcal{N}}$, define $G_a \times (\mathbb{R}^3)^{\mathcal{N}} \to (\mathbb{R}^3)^{\mathcal{N}} : (g, R) \mapsto (i \mapsto g \cdot \mathbf{R}_i)$; that is, $[(\mathbf{b}, A)R]_i = \mathbf{b} + A\mathbf{R}_i$ for all $i \in \mathcal{N}$. Then G_a also acts on $(\mathbb{R}^3)^{\mathcal{N}}$ from the left by this mapping.

We will illustrate this concept with a rigid motion of a water molecule, H_2O. Let $\mathcal{N} = \{O, H_1, H_2\}$. We take the bond length between the oxygen atom and each hydrogen atom to be 1 angstrom and the angle between the two bonds to be 104.5°. So the position vectors of the atoms are $\mathbf{R}_O = \left(\begin{smallmatrix} 0 \\ 0 \\ 0 \end{smallmatrix}\right), \mathbf{R}_{H_1} = \left(\begin{smallmatrix} 1 \\ 0 \\ 0 \end{smallmatrix}\right)$ and $\mathbf{R}_{H_2} = \left(\begin{smallmatrix} \cos 104.5^\circ \\ \sin 104.5^\circ \\ 0 \end{smallmatrix}\right)$ for O, H_1, H_2, respectively. This is a particular configuration of the molecule. Suppose we would like to move the molecule up one unit in the z-axis direction and rotate it 90° about the z-axis in the positive direction by the right-handed rule. Then the translation vector would be $\mathbf{b} = \left(\begin{smallmatrix} 0 \\ 0 \\ 1 \end{smallmatrix}\right)$, and the rotation matrix would be $A = \left(\begin{smallmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)$. So we let $(\mathbf{b}, A) = \left(\left(\begin{smallmatrix} 0 \\ 0 \\ 1 \end{smallmatrix}\right), \left(\begin{smallmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)\right) \in G_a$ be applied to R. If R' denotes the new configuration, then $R' = (\mathbf{b}, A)R$, and

$$\begin{aligned}
\mathbf{R}'_O = (\mathbf{b}, A) \cdot \mathbf{R}'_O &= \left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right) \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.
\end{aligned}$$

Similarly,

$$\mathbf{R}'_{H_1} = (\mathbf{b}, A) \cdot \mathbf{R}_{H_1} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{R}'_{H_2} = (\mathbf{b}, A) \cdot \mathbf{R}_{H_2} = \begin{pmatrix} -\sin 104.5^\circ \\ \cos 104.5^\circ \\ 1 \end{pmatrix}.$$

If a group G acts on a set X from the left, then the set $Gx = \{gx \in X \mid g \in G\}$, where $x \in X$, is called an **orbit**, and x is called a **representative** of the orbit. Also $G\backslash\!\backslash X := \{Gx \mid x \in X\}$ is the set of all the orbits of this action. In our case, G_aR is an orbit where the group is G_a and the representative configuration is R, and the orbit G_aR is called a **conformation**.

1.2. Definitions and Notations

Traditionally, a graph $G = (V, E)$ is a finite collection V of objects called **vertices** together with a collection E of two element subsets of V called **edges**, and where the traditional incidence relation is defined as follows: a vertex $v \in V$ is **incident** on an edge $e \in E$ if and only if v is an element of e.

Definition. An **abstract graph** is defined as a triple (V, E, ι) where V is a set of vertices, E is another set, the elements of which are called edges, and $\iota \subset V \times E$ is an incidence relation between members of V and members of E, such that the following conditions hold:

(1) for every edge $e \in E$, there are exactly two vertices, $v, w \in V$ such that $(v, e), (w, e) \in \iota$; and

(2) for every pair of distinct vertices, $v, w \in V$, there is at most one $e \in E$ such that $(v, e), (w, e) \in \iota$.

Note in this definition, an edge may not be a two-element subset of V. A graph with the traditional definition of its incidence relation is a particular example of an abstract graph. We will call it a **traditional graph** to distinguish it from other abstract graphs whose edges and incidence relations are defined differently. Let S be any set and let $\mathcal{P}(S)$ denote the set of all subsets of S. We define a **graph** $G = (V, E)$, for $V \subset \mathcal{P}(S)$ and $E \subset \mathcal{P}(S)$, to be with the incidence relation defined as $v \in V$ is **incident** on $e \in E$ if and only if v is a subset of e, and where the two conditions of an abstract graph hold true.

From Figure 1, some examples of the incidence relation are:

- u_3 is incident on e_2, u_4 and u_8 on e_7, e_1 on u_1, and e_5 on u_3 in (a).
- v_5 is incident on f_5, f_2 on v_1 and v_3, and f_6 on v_4 in (b).
- w_3 is incident on g_2 and g_2 on w_2 in (c). No edge is incident on w_5.

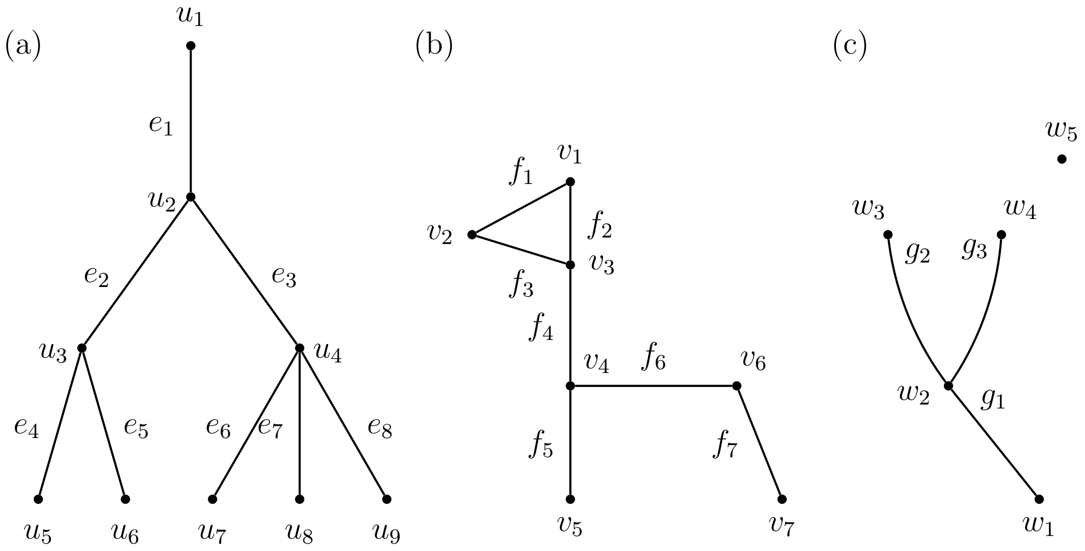


Figure 1. Examples of Graph

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
u_1	1	0	0	0	0	0	0	0
u_2	1	1	1	0	0	0	0	0
u_3	0	1	0	1	1	0	0	0
u_4	0	0	1	0	0	1	1	1
u_5	0	0	0	1	0	0	0	0
u_6	0	0	0	0	1	0	0	0
u_7	0	0	0	0	0	1	0	0
u_8	0	0	0	0	0	0	1	0
u_9	0	0	0	0	0	0	0	1

Table 1. Example of an Incidence Table

This relationship can be shown as a table as well. For example, Table 1 is the incidence table for the graph (a) of Figure 1. In Table 1, a 1 in row i and column j indicates that the vertex u_i is incident on the edge e_j and therefore $(u_i, e_j) \in \iota$.

A vertex v is said to be a **leaf** if there is only one edge e incident on v. So if the row of a vertex has only one 1 in the incidence table, then the vertex is a leaf. There are exactly two 1's in each column; this is due to the first condition in the definition of an abstract graph.

In a graph $G = (V, E)$, let p be an alternating sequence of vertices and edges, $p = (v_0, e_1, v_1, e_2, v_2, \ldots, v_{m-1}, e_m, v_m)$ where $e_j \in E$ is incident on both $v_{j-1} \in V$ and $v_j \in V$ for $j \in \{1, 2, \ldots, m\}$. Also let $v_1, \ldots, v_m$ be distinct vertices and $e_1, \ldots, e_m$ be distinct edges. If v_0 is distinct from $v_1, \ldots, v_m$, then p is called a **path of length** m from v_0 to v_m. If we have a path connecting any two distinct vertices in the graph, we say that the graph is **connected**. If $v_0 = v_m$ and $m \geq 3$, p is called a **cycle**; a graph which does not have any cycle is said to be **acyclic**. A graph which is acyclic and connected is called a **tree**. From the examples in Figure 1, the graph (a) is the only tree out of three examples since the graph (b) is connected but has a cycle $(v_1, f_1, v_2, f_2, v_3, f_3, v_1)$, and the graph (c) is acyclic but not connected for no path exists to w_5. A **rooted graph** (V, E, v) where $v \in V$ is a graph (V, E) with a vertex chosen which is called the **root vertex**.

For any set S, we define $\binom{S}{r}$ to be the set which contains all the r-element subsets of S, subsets of S with exactly r elements. Let $\mathcal{N}$ be a set of N elements. Then an **abstract** k**-simplex** is an element of $\binom{\mathcal{N}}{k+1}$ for $k < N$; an element of $\binom{\mathbb{R}^3}{k+1}$ is called a k**-simplex**. For an abstract k-simplex $s = \{i_0, i_1, \ldots, i_k\}$, if $R \in (\mathbb{R}^3)^{\mathcal{N}}$, then $R_s = \{\mathbf{R}_{i_0}, \mathbf{R}_{i_1}, \ldots, \mathbf{R}_{i_k}\}$ is the **associated simplex**. There are 24 possible orderings of the four-element set, $\{0, 1, 2, 3\}$; that is, $|S_4| = 24$. If $n \geq 2$, then every permutation in S_n can be written as a product of transpositions. The number of factors occurring in any factorization of $\pi \in S_n$ into transpositions is either always even or always odd [26]. Those that factor into an even number of transpositions are called **even permutations**, and the others are called **odd permutations**. Define **oriented** 3**-simplices** to be equivalence classes

$$[A_0, A_1, A_2, A_3] = \{(A_{\pi(0)}, \ldots, A_{\pi(3)}) \mid \pi \text{ is an even permutation of } \{0, 1, 2, 3\}\},$$

where $\{A_0, A_1, A_2, A_3\}$ is an abstract 3-simplex. For example, since $\pi = (021) = (01)(02)$ for $(2, 0, 1, 3) = \pi(0, 1, 2, 3)$, (A_0, A_1, A_2, A_3) is equivalent to (A_2, A_0, A_1, A_3).

However, it is not equivalent to (A_3, A_0, A_1, A_2) since in this case, the permutation is $\pi = (0321) = (12)(01)(03)$ which is odd. (A_3, A_0, A_1, A_2) is equivalent to (A_3, A_1, A_2, A_0). With this definition, we get exactly two equivalence classes: one class with orders resulting from applying even permutations to the indices of (A_0, A_1, A_2, A_3) and another with orders obtained by applying odd permutations. The abstract simplex $\{A_0, A_1, A_2, A_3\}$ is the **underlying 3-simplex** for either orientation, $[A_0, A_1, A_2, A_3]$ or $[A_3, A_1, A_2, A_0]$, and we can define a mapping $d^* \mapsto d$ from an oriented 3-simplex d^* to its underlying 3-simplex d. For $0 \leq k \leq 3$, a simplex $\{\mathbf{R}_0, \mathbf{R}_1, \ldots, \mathbf{R}_k\}$ is said to be **geometrically independent** if the set

$$\left\{ \begin{pmatrix} 1 \\ \mathbf{R}_0 \end{pmatrix}, \begin{pmatrix} 1 \\ \mathbf{R}_1 \end{pmatrix}, \ldots, \begin{pmatrix} 1 \\ \mathbf{R}_k \end{pmatrix} \right\}, \text{ where } \begin{pmatrix} 1 \\ \mathbf{R}_l \end{pmatrix} = \begin{pmatrix} 1 \\ a \\ b \\ c \end{pmatrix} \text{ if } \mathbf{R}_l = \begin{pmatrix} a \\ b \\ c \end{pmatrix},$$

is linearly independent in $\mathbb{R}^4$. Note that any subset of a geometrically independent set is also geometrically independent.

A set $\{s_0, s_1, \ldots, s_k\}$ for $k = 0, 1, 2$, is called a k**-site** if $s_0 \subset s_1 \subset \cdots \subset s_k \subset \mathcal{N}$ for $\mathcal{N}$, a set of N elements, with $|s_0| = 1, |s_1| = 2, \ldots, |s_k| = k + 1$. We can also think of this as an ordered $k + 1$ tuple $r = (i_0, i_1, \ldots, i_k)$ of distinct elements for $s_0(r) = \{i_0\}, s_1(r) = \{i_0, i_1\}, \ldots, s_k(r) = \{i_0, i_1, \ldots, i_k\}$. The set of abstract simplices $\{s_0(r), \ldots, s_k(r)\}$ is called the **flag** associated to the site r. For $k = 0, 1, 2$, a k**-pose** is a 3×2^k real matrix $(\mathbf{e}_0, \mathbf{e}_1, \ldots \mathbf{e}_{2^k - 1})$ such that $\mathbf{e}_0$ is some vector which gives a point in space if $k \geq 0$, $\mathbf{e}_1$ is a unit vector giving a direction in space if $k \geq 1$, and $\mathbf{e}_2, \mathbf{e}_3$ are unit vectors such that $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) \in \mathrm{SO}(3)$, forming a positively oriented orthonormal basis when $k = 2$. Given a k-site $r = (i_0, i_1, \ldots, i_k)$ for $k = 0, 1, 2$, and a configuration $R \in (\mathbb{R}^3)^{\mathcal{N}}$ such that the associated simplex $R_{s_k(r)} = \{\mathbf{R}_{i_0}, \mathbf{R}_{i_1}, \ldots, \mathbf{R}_{i_k}\}$ is geometrically independent, the k**-pose at** r **conformed to** R is the 3×2^k matrix

$E_r(R) = (\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_{2^k-1})$ whose column vectors are:

$$\begin{aligned}
\mathbf{e}_0 &= \mathbf{R}_{i_0} \qquad \text{if } k \geq 0, \\
\mathbf{e}_1 &= \frac{\mathbf{R}_{i_1} - \mathbf{R}_{i_0}}{\|\mathbf{R}_{i_1} - \mathbf{R}_{i_0}\|} \qquad \text{if } k \geq 1, \\
\mathbf{e}_2 &= \frac{(I - \mathbf{e}_1\mathbf{e}_1^T)(\mathbf{R}_{i_2} - \mathbf{R}_{i_0})}{\|(I - \mathbf{e}_1\mathbf{e}_1^T)(\mathbf{R}_{i_2} - \mathbf{R}_{i_0})\|} \qquad \text{and} \\
\mathbf{e}_3 &= \mathbf{e}_1 \times \mathbf{e}_2 \qquad \text{if } k = 2.
\end{aligned}$$

It is easy to see that $\mathbf{e}_0$ and $\mathbf{e}_1$ are well-defined since $\{\mathbf{R}_{i_0}, \mathbf{R}_{i_1}, \ldots, \mathbf{R}_{i_k}\}$ is geometrically independent. $\mathbf{e}_2$ is also well-defined. Suppose not; so assume $\|(I - \mathbf{e}_1\mathbf{e}_1^T)(\mathbf{R}_{i_2} - \mathbf{R}_{i_0})\| = 0$. Then $\mathbf{R}_{i_2} - \mathbf{R}_{i_0} = \mathbf{e}_1\mathbf{e}_1^T(\mathbf{R}_{i_2} - \mathbf{R}_{i_0})$, implying that $\mathbf{R}_{i_2} - \mathbf{R}_{i_0} = c\,\mathbf{e}_1$ for $c = \mathbf{e}_1^T(\mathbf{R}_{i_2} - \mathbf{R}_{i_0})$. So $\mathbf{R}_{i_2} - \mathbf{R}_{i_0} = c\,\frac{\mathbf{R}_{i_1} - \mathbf{R}_{i_0}}{\|\mathbf{R}_{i_1} - \mathbf{R}_{i_0}\|}$, which implies that $\{\mathbf{R}_{i_0}, \mathbf{R}_{i_1}, \mathbf{R}_{i_2}\}$ is collinear, resulting in a contradiction. Thus, $\mathbf{e}_2$ is well-defined, and so consequently, $\mathbf{e}_3$ is as well. We denote the set of all 2-poses by $\mathcal{P}$.

Fact. *If the k-site r, $k = 0, 1, 2$, and the configuration R are such that $R_{s_k(r)}$ is geometrically independent, then for all $(\mathbf{b}, A) \in G_a$, $[(\mathbf{b}, A)R]_{s_k(r)}$ is also geometrically independent and $(\mathbf{b}, A) \cdot E_r(R) = E_r((\mathbf{b}, A)R)$.*

Proof. Since $R_{s_k(r)}$ is geometrically independent, for the case of $k = 2$, we have $\left\{\binom{1}{\mathbf{R}_{i_0}}, \binom{1}{\mathbf{R}_{i_1}}, \binom{1}{\mathbf{R}_{i_2}}\right\}$ is linearly independent in $\mathbb{R}^4$ for $r = (i_0, i_1, i_2)$. Recall that $[(\mathbf{b}, A)R]_i = \mathbf{b} + A\mathbf{R}_i$ for all $i \in \mathcal{N}$. In particular, $(\mathbf{b}, A)R$ maps $\mathbf{R}_{i_0}$ to $\mathbf{b} + A\mathbf{R}_{i_0}$, $\mathbf{R}_{i_1}$ to $\mathbf{b} + A\mathbf{R}_{i_1}$, and $\mathbf{R}_{i_2}$ to $\mathbf{b} + A\mathbf{R}_{i_2}$. So we need to show that $\left\{\binom{1}{\mathbf{b}+A\mathbf{R}_{i_0}}, \binom{1}{\mathbf{b}+A\mathbf{R}_{i_1}}, \binom{1}{\mathbf{b}+A\mathbf{R}_{i_2}}\right\}$ is linearly independent in $\mathbb{R}^4$. So let

$$\alpha_0 \begin{pmatrix} 1 \\ \mathbf{b} + A\mathbf{R}_{i_0} \end{pmatrix} + \alpha_1 \begin{pmatrix} 1 \\ \mathbf{b} + A\mathbf{R}_{i_1} \end{pmatrix} + \alpha_2 \begin{pmatrix} 1 \\ \mathbf{b} + A\mathbf{R}_{i_2} \end{pmatrix} = 0$$

for some constants $\alpha_0, \alpha_1, \alpha_2$. Then

$$\begin{pmatrix} \alpha_0 \\ \alpha_0(\mathbf{b} + A\mathbf{R}_{i_0}) \end{pmatrix} + \begin{pmatrix} \alpha_1 \\ \alpha_1(\mathbf{b} + A\mathbf{R}_{i_1}) \end{pmatrix} + \begin{pmatrix} \alpha_2 \\ \alpha_2(\mathbf{b} + A\mathbf{R}_{i_2}) \end{pmatrix} = 0.$$

Thus, $\alpha_0 + \alpha_1 + \alpha_2 = 0$ and $\alpha_0(\mathbf{b} + A\mathbf{R}_{i_0}) + \alpha_1(\mathbf{b} + A\mathbf{R}_{i_1}) + \alpha_2(\mathbf{b} + A\mathbf{R}_{i_2}) = 0$. Then we get $(\alpha_0 + \alpha_1 + \alpha_2)\mathbf{b} + A(\alpha_0\mathbf{R}_{i_0} + \alpha_1\mathbf{R}_{i_1} + \alpha_2\mathbf{R}_{i_2}) = 0$. Since $A \in \mathrm{SO}(3)$ is invertible, with $\alpha_0 + \alpha_1 + \alpha_2 = 0$, we have $\alpha_0\mathbf{R}_{i_0} + \alpha_1\mathbf{R}_{i_1} + \alpha_2\mathbf{R}_{i_2} = 0$; therefore, $\alpha_0 \left(\begin{smallmatrix} 1 \\ \mathbf{R}_{i_0} \end{smallmatrix}\right) + \alpha_1 \left(\begin{smallmatrix} 1 \\ \mathbf{R}_{i_1} \end{smallmatrix}\right) + \alpha_2 \left(\begin{smallmatrix} 1 \\ \mathbf{R}_{i_2} \end{smallmatrix}\right) = 0$. Because $\left\{\left(\begin{smallmatrix} 1 \\ \mathbf{R}_{i_0} \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 \\ \mathbf{R}_{i_1} \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 \\ \mathbf{R}_{i_2} \end{smallmatrix}\right)\right\}$ is linearly independent, we have $\alpha_0 = \alpha_1 = \alpha_2 = 0$. Therefore, $\left\{\left(\begin{smallmatrix} 1 \\ \mathbf{b}+A\mathbf{R}_{i_0} \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 \\ \mathbf{b}+A\mathbf{R}_{i_1} \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 \\ \mathbf{b}+A\mathbf{R}_{i_2} \end{smallmatrix}\right)\right\}$ is linearly independent in $\mathbb{R}^4$. Hence, $[(\mathbf{b}, A)R]_{s_2(r)}$ is geometrically independent, and we conclude that $[(\mathbf{b}, A)R]_{s_k(r)}$ is geometrically independent for $k = 2$. For $k = 1$, since $R_{s_1(r)}$ is geometrically independent, we have $\left\{\left(\begin{smallmatrix} 1 \\ \mathbf{R}_{i_0} \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 \\ \mathbf{R}_{i_1} \end{smallmatrix}\right)\right\}$ is linearly independent in $\mathbb{R}^4$ for $r = (i_0, i_1)$. Then for some constants α_0, α_1, let

$$\alpha_0 \begin{pmatrix} 1 \\ \mathbf{b} + A\mathbf{R}_{i_0} \end{pmatrix} + \alpha_1 \begin{pmatrix} 1 \\ \mathbf{b} + A\mathbf{R}_{i_1} \end{pmatrix} = 0.$$

Then $\alpha_0 + \alpha_1 = 0$ and $\alpha_0(\mathbf{b} + A\mathbf{R}_{i_0}) + \alpha_1(\mathbf{b} + A\mathbf{R}_{i_1})) = 0$ so that $(\alpha_0 + \alpha_1)\mathbf{b} + A(\alpha_0\mathbf{R}_{i_0} + \alpha_1\mathbf{R}_{i_1}) = 0$. By the similar argument as in the case of $k = 2$, we have $\alpha_0 = \alpha_1 = 0$. Thus, we conclude that $\left\{\left(\begin{smallmatrix} 1 \\ \mathbf{b}+A\mathbf{R}_{i_0} \end{smallmatrix}\right), \left(\begin{smallmatrix} 1 \\ \mathbf{b}+A\mathbf{R}_{i_1} \end{smallmatrix}\right)\right\}$ is linearly independent in $\mathbb{R}^4$; and therefore, $[(\mathbf{b}, A)R]_{s_k(r)}$ is geometrically independent for $k = 1$. The case when $k = 0$ is trivial. Hence, we have $[(\mathbf{b}, A)R]_{s_k(r)}$ is geometrically independent for $k = 0, 1, 2$.

Since $[(\mathbf{b}, A)R]_{s_k(r)}$ is geometrically independent, $E_r((\mathbf{b}, A)R)$ is well-defined. We also know that $(\mathbf{b}, A) \cdot E_r(R) = (\mathbf{b} + A\mathbf{e}_0, A\mathbf{e}_1, A\mathbf{e}_2, A\mathbf{e}_3)$. So for $E_r((\mathbf{b}, A)R) = (\mathbf{e}'_0, \mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$, it is easy to see that $\mathbf{e}'_0 = \mathbf{b} + A\mathbf{R}_{i_0} = \mathbf{b} + A\mathbf{e}_0$. For $\mathbf{e}'_1$, we have:

$$\begin{aligned} \mathbf{e}'_1 &= \frac{\mathbf{b} + A\mathbf{R}_{i_1} - (\mathbf{b} + A\mathbf{R}_{i_0})}{\|\mathbf{b} + A\mathbf{R}_{i_1} - (\mathbf{b} + A\mathbf{R}_{i_0})\|} = \frac{A\mathbf{R}_{i_1} - A\mathbf{R}_{i_0}}{\|A\mathbf{R}_{i_1} - A\mathbf{R}_{i_0}\|} \\ &= \frac{A(\mathbf{R}_{i_1} - \mathbf{R}_{i_0})}{\|A(\mathbf{R}_{i_1} - \mathbf{R}_{i_0})\|} = A\left(\frac{\mathbf{R}_{i_1} - \mathbf{R}_{i_0}}{\|\mathbf{R}_{i_1} - \mathbf{R}_{i_0}\|}\right) \\ &= A\mathbf{e}_1. \end{aligned}$$

Note that we have $\|A(\mathbf{R}_{i_1} - \mathbf{R}_{i_0})\| = \|\mathbf{R}_{i_1} - \mathbf{R}_{i_0}\|$ since $A \in \mathrm{SO}(3)$. SInce $\mathbf{e}'_1 = A\mathbf{e}_1$, $\mathbf{e}'_1(\mathbf{e}'_1)^T = (A\mathbf{e}_1)(A\mathbf{e}_1)^T = A\mathbf{e}_1\mathbf{e}_1^T A^T$; thus, $I - \mathbf{e}'_1(\mathbf{e}'_1)^T = AA^T - A\mathbf{e}_1\mathbf{e}_1^T A^T = A(I -$

$\mathbf{e}_1\mathbf{e}_1^T)A^T$. So for $\mathbf{e}_2'$, since $A^TA = I$,

$$\begin{aligned}\mathbf{e}_2' &= \frac{[A(I-\mathbf{e}_1\mathbf{e}_1^T)A^T][A(\mathbf{R}_{i_2}-\mathbf{R}_{i_0})]}{\|[A(I-\mathbf{e}_1\mathbf{e}_1^T)A^T][A(\mathbf{R}_{i_2}-\mathbf{R}_{i_0})]\|} \\ &= A\left[\frac{(I-\mathbf{e}_1\mathbf{e}_1^T)(\mathbf{R}_{i_2}-\mathbf{R}_{i_0})}{\|(I-\mathbf{e}_1\mathbf{e}_1^T)(\mathbf{R}_{i_2}-\mathbf{R}_{i_0})\|}\right] \\ &= A\mathbf{e}_2.\end{aligned}$$

Then, we have $\mathbf{e}_3' = \mathbf{e}_1' \times \mathbf{e}_2' \ = A\mathbf{e}_1 \times A\mathbf{e}_2 = A(\mathbf{e}_1 \times \mathbf{e}_2) = A\mathbf{e}_3$. Therefore, $(\mathbf{b}, A) \cdot E_r(R) = E_r((\mathbf{b}, A)R)$. □

Also for $\left(\begin{smallmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b}' & A' \end{smallmatrix}\right) \in G_p$ with $E_r(R) \in \mathcal{P}$ and $(\mathbf{b}, A) \in G_a$, the commutative property of G_a and G_p from Section 1.1 implies $[(\mathbf{b}, A) \cdot E_r(R)]\left(\begin{smallmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b}' & A' \end{smallmatrix}\right) = (\mathbf{b}, A) \cdot \left[E_r(R)\left(\begin{smallmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b}' & A' \end{smallmatrix}\right)\right]$.

Define:

$$\begin{aligned} G_p^{(2)} = &\quad G_p = \left\{ \begin{pmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b} & A \end{pmatrix} \middle| \mathbf{b} \in \mathbb{R}^3, A \in \mathrm{SO}(3) \right\}; \\ G_p^{(1)} = &\quad \left\{ \begin{pmatrix} 1 & 0 \\ \mathbf{b} & \mathbf{u} \end{pmatrix} \middle| \mathbf{b} \in \mathbb{R}^3, \mathbf{u} \in S^2 \right\} \quad \begin{array}{l} \text{where } S^2 = \{\mathbf{u} \in \mathbb{R}^3 \mid \|\mathbf{u}\| = 1\} \\ \text{is the two dimensional sphere in } \mathbb{R}^3 \end{array}; \\ G_p^{(0)} = &\quad \left\{ \begin{pmatrix} 1 \\ \mathbf{b} \end{pmatrix} \middle| \mathbf{b} \in \mathbb{R}^3 \right\}. \end{aligned}$$

Fact. *Suppose $E = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ is a 2-pose and $E' = (\mathbf{e}_0', \ldots, \mathbf{e}_{2^k-1})$ is a k-pose. Then there exists a unique $\mathcal{A} \in G_p^{(k)}$ such that $E' = E\mathcal{A}$ for $k = 0, 1, 2$.*

Proof. To see the existence and uniqueness of the matrix, suppose

$$E = (\mathbf{b}_1, A_1), \quad E' = \begin{cases} (\mathbf{b}_2, A_2) & \text{if } k = 2 \\ (\mathbf{b}_2, \mathbf{u}_2) & \text{if } k = 1 \\ (\mathbf{b}_2) & \text{if } k = 0 \end{cases}, \quad \text{and} \quad \mathcal{A} = \begin{cases} \left(\begin{smallmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b} & A \end{smallmatrix}\right) & \text{if } k = 2 \\ \left(\begin{smallmatrix} 1 & 0 \\ \mathbf{b} & \mathbf{u} \end{smallmatrix}\right) & \text{if } k = 1 \\ \left(\begin{smallmatrix} 1 \\ \mathbf{b} \end{smallmatrix}\right) & \text{if } k = 0 \end{cases}.$$

Then,

$$E\mathcal{A} = \begin{cases} (\mathbf{b}_1 + A_1\mathbf{b}, A_1A) & \text{for } k = 2 \\ (\mathbf{b}_1 + A_1\mathbf{b}, A_1\mathbf{u}) & \text{for } k = 1 \\ (\mathbf{b}_1 + A_1\mathbf{b}) & \text{for } k = 0 \end{cases};$$

therefore,

$$\begin{cases} \mathbf{b} = A_1^T(\mathbf{b}_2 - \mathbf{b}_1) & \text{for } k = 0, 1, 2 \\ \mathbf{u} = A_1^T\mathbf{u}_2 & \text{for } k = 1 \\ A = A_1^T A_2 & \text{for } k = 2 \end{cases}.$$

□

Given a 2-site r, a k-site r', and a configuration R, for which $R_{s_2(r)}$ and $R_{s_k(r')}$ are geometrically independent, there exists a unique 4×2^k matrix $\mathcal{A}_{r,r'}(R)$ in $G_p^{(k)}$ such that $E_{r'}(R) = E_r(R)\mathcal{A}_{r,r'}(R)$. Such a matrix is called a **coordinate transformation matrix**.

For a 2-site $r = (i_0, i_1, i_2)$ with its pose $E_r(R) = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and a k-site $r' = (i'_0, \dots, i'_k)$ and its pose $E_{r'}(R) = (\mathbf{e}'_0, \dots, \mathbf{e}'_{2^k-1})$ for $k = 0, 1, 2$, the vector $\mathbf{b} = (x, y, z)^T$ of $\mathcal{A}_{r,r'} = \left(\begin{smallmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b} & A \end{smallmatrix}\right) \in G_p^{(k)}$ tells us the position of i'_0 relative to $\mathbf{e}_0$, the position of i_0, using the basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, so, $\mathbf{e}'_0 = \mathbf{e}_0 + x\mathbf{e}_1 + y\mathbf{e}_2 + z\mathbf{e}_3$. When $k = 1$, the vector $\mathbf{u} = (x', y', z')^T$ in $\mathcal{A}_{r,r'} = \left(\begin{smallmatrix} 1 & 0 \\ \mathbf{b} & \mathbf{u} \end{smallmatrix}\right) \in G_p^1$ is the unit vector which gives the direction from i'_0 to i'_1 in terms of the basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ by $\mathbf{e}'_1 = x'\mathbf{e}_0 + y'\mathbf{e}_1 + z'\mathbf{e}_2$. When $k = 2$, the matrix A in $\mathcal{A}_{r,r'} = \left(\begin{smallmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b} & A \end{smallmatrix}\right) \in G_p^{(2)}$ tells us the vectors in the orthonormal basis $(\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$ in terms of the orthonormal basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. So by applying $\mathcal{A}_{r,r'}$ to the pose $E_r(R)$, we transform the coordinate system based on $E_r(R)$ to the coordinate system based on $E_{r'}(R)$. We will see later in Section 1.5 that $\mathcal{A}_{r,r'}$ does not depend strongly on the particular configuration.

The subscript a in G_a stands for **active**, and p in G_p is for **passive**. G_a is active in the sense that by acting on a pose from the left, an element of G_a moves the pose at a site in a particular configuration to the pose at the same site in another configuration; so, the action carries the sense of moving the molecule from one configuration to the other. On the other hand, an element of the passive group G_p would transform the pose at a site of a configuration to the pose of another site on the same configuration. Using the fact that $E_{r'}(R) = E_r(R)\mathcal{A}_{r,r'}(R)$ for two 2-sites r and r', we can obtain another relationship useful in geometry calculations. Let $\mathcal{A}_{r,r'}(R) = \left(\begin{smallmatrix} 1 & \boldsymbol{\theta}^T \\ \mathbf{b} & A \end{smallmatrix}\right)$. Also given a position vector $\mathbf{v}$ of an atom in the reference coordinate system, let $\mathbf{x} = \left(\begin{smallmatrix} 1 \\ \mathbf{w} \end{smallmatrix}\right)$ and $\mathbf{x}' = \left(\begin{smallmatrix} 1 \\ \mathbf{w}' \end{smallmatrix}\right)$ be such that $\mathbf{v} = E_r(R)\mathbf{x}$ and $\mathbf{v} = E_{r'}(R)\mathbf{x}'$. So the vectors $\mathbf{w}$ and $\mathbf{w}'$ multiplied by the orthonormal bases of the poses $E_r(R)$ and $E_{r'}(R)$, repectively, give the position of $\mathbf{v}$ in terms of the respective bases. Then by the fact that $E_{r'}(R) = E_r(R)\mathcal{A}_{r,r'}(R)$, we have $E_r(R)\mathbf{x} = E_r(R)\mathcal{A}_{r,r'}(R)\mathbf{x}'$. Thus $\mathbf{x} = \mathcal{A}_{r,r'}(R)\mathbf{x}'$, that is,

$$\begin{pmatrix} 1 \\ \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{\theta}^T \\ \\ \mathbf{b} & A \end{pmatrix} \begin{pmatrix} 1 \\ \\ \mathbf{w}' \end{pmatrix} = \begin{pmatrix} 1 \\ \\ \mathbf{b} + A\mathbf{w}' \end{pmatrix}.$$

So the relationship between the vectors $\mathbf{w}$ and $\mathbf{w}'$ is described by $\mathbf{w} = \mathbf{b} + A\mathbf{w}'$.

1.3. The Definition of Generalized Z-systems

Let $\Gamma \subset \mathcal{P}(\mathcal{N})$. Let $\Gamma^k \subset \Gamma$ denote the set of all abstract k-simplices in Γ. If $\mathcal{C} \subset \mathcal{N}$, then let $\Gamma_{\mathcal{C}} \subset \Gamma$ denote the set of all simplices that are subsets of $\mathcal{C}$. The set of all k-simplices which are subsets of $\mathcal{C}$ and are in Γ is denoted by $\Gamma^k_{\mathcal{C}} = \Gamma^k \cap \Gamma_{\mathcal{C}}$.

Definition. $\Gamma \subset \mathcal{P}(\mathcal{N})$ for a set $\mathcal{N}$ is called an **unoriented generalized Z-system** (**GZ-system**) if the following conditions hold:

(1) if $e \in \Gamma$, then $1 \leq |e| \leq 4$;

(2) $\Gamma^0 = \binom{\mathcal{N}}{1}$.

(3) if $e \in \Gamma^k$ for $1 \leq k \leq 3$, then $|\Gamma^{k-1}_e| = 2$;

(4) (Γ^0, Γ^1) is an acyclic graph;

(5) if $\mathcal{C} \subset \mathcal{N}$ is such that $(\Gamma^0_{\mathcal{C}}, \Gamma^1_{\mathcal{C}})$ is a connected component of (Γ^0, Γ^1), then $(\Gamma^1_{\mathcal{C}}, \Gamma^2_{\mathcal{C}})$ and $(\Gamma^2_{\mathcal{C}}, \Gamma^3_{\mathcal{C}})$ are (possibly empty) trees (Such a subset $\mathcal{C}$ is called a **component**);

(6) $\Gamma^2 \neq \emptyset$; and

(7) if $v_1, v_2 \in \Gamma^2$ and $v_1 \cup v_2 \in \Gamma^3$, then $v_1 \cap v_2 \in \Gamma^1$.

If (Γ^0, Γ^1) is also connected, that is to say (Γ^0, Γ^1) is a tree, then we say Γ is an **unoriented Z-system**.

Let Γ^3_* denote a set of oriented 3-simplices such that the mapping $d^* \mapsto d$ from an oriented 3-simplex $d^* \in \Gamma^3_*$ to its underlying 3-simplex $d \in \Gamma^3$ is a bijection. Then $\Gamma^* = (\Gamma^1, \Gamma^2, \Gamma^3_*)$ is an **oriented (G)Z-system** or simply a **(G)Z-system**.

From the third condition in the above definition, we may think of $e \in \Gamma^k$ for $k = 1, 2, 3$, as an edge incident on the two vertices $v_1, v_2 \in \Gamma^{k-1}$ where $\Gamma^{k-1}_e = \{v_1, v_2\}$. By this condition, the first of the two conditions for abstract graph—of an edge being incident on exactly two vertices—is covered. Note that in regard to conditions (4) and (5) of the above definition, in each graph $G = (V, E)$, V is a collection of k element subsets of $\mathcal{N}$ for $k \geq 1$, and E is a collection of $k + 1$ element subsets of

$\mathcal{N}$. By condition (3), each $e \in E$ is incident on $v_1, v_2 \in V$, $v_1 \neq v_2$, and both v_1 and v_2 have k elements. Therefore $v_1 \cup v_2 \subset e$ has more than k elements, and since we know that e has only $k+1$ elements, we have $e = v_1 \cup v_2$. So the second condition for an abstract graph—an edge being uniquely determined by the vertices on which it is incident—follows automatically. The sixth condition insures that at least one component of the system has more than two atoms in it.

In the study of the geometry of a biomolecular system, we let $\mathcal{N}$ be the set of atom names of all the atoms in the system, and the elements of $\Gamma^0, \Gamma^1, \Gamma^2$, and Γ^3 are called **atoms, bonds, triangles** and **tetrahedra** respectively. For $b_0, b_1 \in \Gamma^1$, $\alpha = \{b_0, b_1\}$ is called an **angle** if $b_0 \cup b_1 \in \Gamma^2$ and $b_0 \cap b_1 \in \Gamma^0$; $\{A\} = b_0 \cap b_1$ is called the **common atom** of α or of the triangle $t = b_0 \cup b_1$. For $t_0, t_1 \in \Gamma^2$, $\omega = \{t_0, t_1\}$ is called a **wedge** if $t_0 \cup t_1 \in \Gamma^3$ and $t_0 \cap t_1 \in \Gamma^1$; $b = t_0 \cap t_1$ is called the **common bond** of ω or of the tetrahedron $d = t_0 \cup t_1$. $(\Gamma^0, \Gamma^1), (\Gamma^1, \Gamma^2)$, and (Γ^2, Γ^3) are called the atom/bond graph, bond/angle graph, and angle/wedge graph respectively.

In an unoriented GZ-system Γ, triangles and angles are in one-to-one correspondence as $t = b_0 \cup b_1 \in \Gamma^2$ for $b_0, b_1 \in \Gamma^1_t$ corresponds to $\alpha = \{b_0, b_1\} = \Gamma^1_t$. The lemma in [8] states that tetrahedra and wedges also are in one-to-one correspondence: $d = t_1 \cup t_2 \in \Gamma^3$ for $t_1, t_2 \in \Gamma^2_d$ corresponds to $\omega = \{t_1, t_2\} = \Gamma^2_d$.

Suppose $\mathcal{C}$ is a component of Γ. If $|\mathcal{C}| = 1$ or $|\mathcal{C}| = 2$, $\mathcal{C}$ is called **monatomic** or **diatomic**, respectively. If $|\mathcal{C}| \geq 3$, $\mathcal{C}$ is **multi-atomic**. So if $\mathcal{N} = \cup_{j=1}^{m} \mathcal{C}_j$ for some $m > 1$ where $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m$ are distinct components of Γ, then Γ is a GZ-system, and Γ is simply a Z-system if $m = 1$; that is, there is only one component. An example of a GZ-system with different types of components is in Figure 2, which will be discussed in detail in Section 1.8.

A GZ-system Γ is **rooted** if the roots of the rooted graphs $(\Gamma^0, \Gamma^1, a), (\Gamma^1, \Gamma^2, b)$, and (Γ^2, Γ^3, t) are chosen so that $\{a, b, t\}$ is a site on a multi-atomic component of Γ.

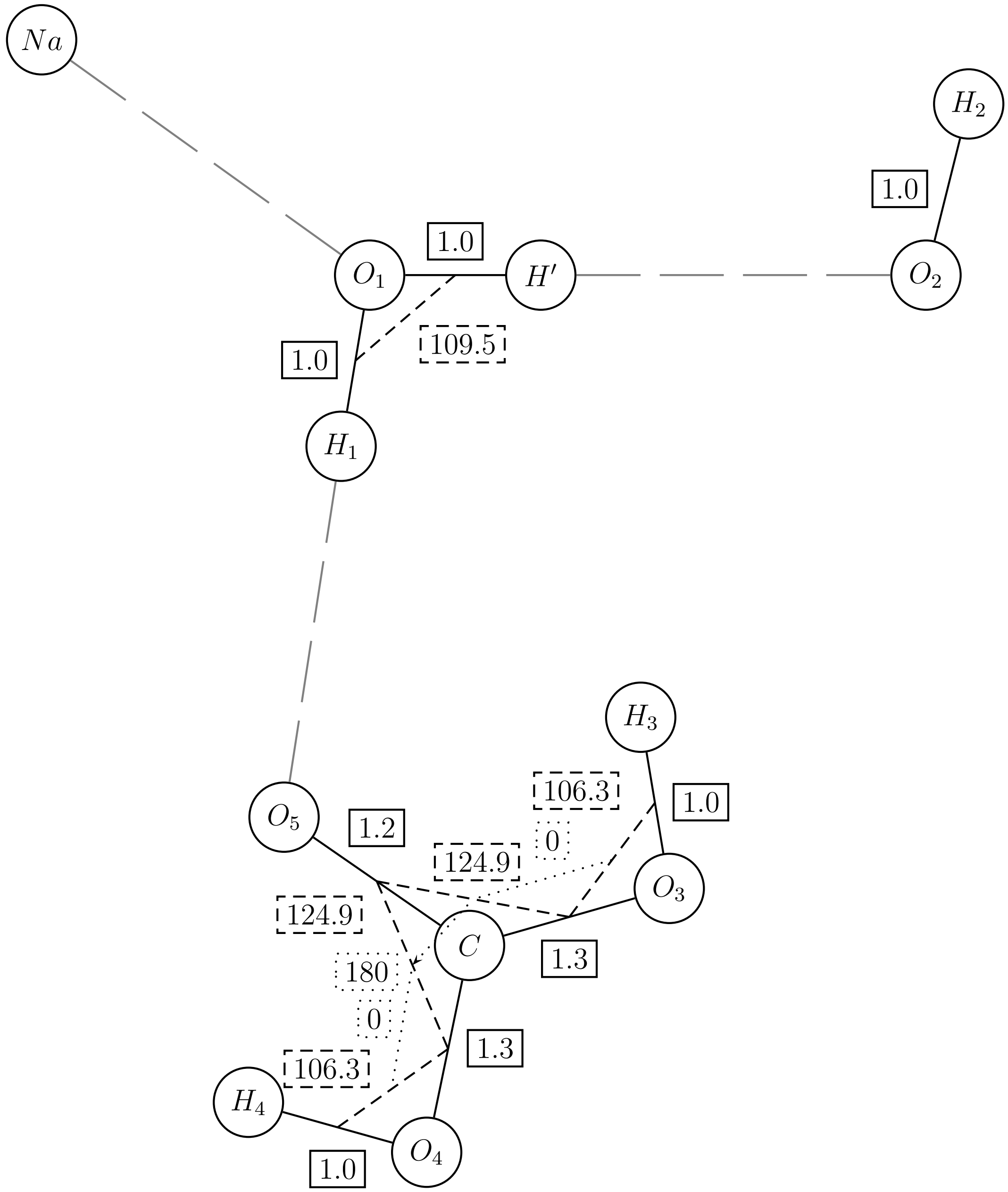


Figure 2. The noncovalent bonds, that are not in the GZ-system, are shown as long gray dashed lines. The circles are atoms labeled with atom names, and (covalent) bonds are shown in solid lines with their lengths given in angstroms in the boxes of solid line close to them. An angle (triangle) is represented by a dashed line between two bonds, and its bond angle is given in degrees, framed with a dashed line. See Section 1.8 for details.

1.4. Conformations and Labeled Z-systems

Define:

$$\mathcal{B}(\Gamma) = \{R \in (\mathbb{R}^3)^{\mathcal{N}} \mid R_s \text{ is geometrically independent for all maximal } s \in \Gamma\backslash\Gamma^3\}$$

for a GZ-system Γ^*. So if $R \in \mathcal{B}(\Gamma)$, then for every bond $b \in \Gamma^1$, the associated 1-simplex R_b will be noncoincident, and also for every triangle $t \in \Gamma^2$, the associated 2-simplex R_t will be noncollinear.

Now let Γ^* be an oriented Z-system and $R \in \mathcal{B}(\Gamma)$ be a particular configuration of the molecular system represented by Γ^*. We will demonstrate that the conformation $G_a R$ can be specified by three mappings: $L : \Gamma^1 \to (0, \infty)$, $C : \Gamma^2 \to (-1, 1)$, and $Z : \Gamma^3_* \to S^1$ where $S^1 = \{z \in \mathbb{C} \mid |z| = 1\}$ is the unit circle in the set of complex numbers. The number $L_b(R)$ to be assigned to every bond $b \in \Gamma^1$ is called a **bond length** and is defined as:

$$L_b(R) = \|\mathbf{R}_{A_0} - \mathbf{R}_{A_1}\|$$

for $b = \{A_0, A_1\} \in \Gamma^1$. Since Γ^* is a Z-system, and therefore is multi-atomic, a maximal element in $\Gamma\backslash\Gamma^3$ is a triangle, and its associated simplex is geometrically independent. Recall that every subset of a geometrically independent set is geometrically independent. Since every bond b of a Z-system Γ is a subset of some triangle $t \in \Gamma^2$, $\{\mathbf{R}_{A_0}, \mathbf{R}_{A_1}\}$ is geometrically independent. Thus, $L_b(R)$, the distance between atoms $\{A_0\}$ and $\{A_1\}$, is always positive. If $\{A_0, A_1\}$ were a diatomic in some GZ-system, then $\{\mathbf{R}_{A_0}, \mathbf{R}_{A_1}\}$ would also be geometrically independent by the definition of $\mathcal{B}(\Gamma)$, so the bond length would also be positive.

Let $t = b_0 \cup b_1 \in \Gamma^2$ for $b_0 = \{A_0, A_1\}, b_1 = \{A_1, A_2\} \in \Gamma^1$. Then define:

$$C_t(R) = \frac{\mathbf{R}_{A_0} - \mathbf{R}_{A_1}}{\|\mathbf{R}_{A_0} - \mathbf{R}_{A_1}\|} \cdot \frac{\mathbf{R}_{A_2} - \mathbf{R}_{A_1}}{\|\mathbf{R}_{A_2} - \mathbf{R}_{A_1}\|}.$$

It gives the cosine of the geometrical angle, called the **bond angle**, between bonds b_0 and b_1. It is always well-defined and $-1 < C_t(R) < 1$ since $\{\mathbf{R}_{A_0}, \mathbf{R}_{A_1}, \mathbf{R}_{A_2}\}$ is noncollinear; in other words, the angle θ between b_0 and b_1 satisfies $0° < \theta < 180°$.

Suppose $t_0 = \{A_0, A_1, A_2\}, t_1 = \{A_1, A_2, A_3\} \in \Gamma^2$, and let an oriented 3-simplex, $d^* = [A_0, A_1, A_2, A_3] \in \Gamma^3_*$ for the underlying 3-simplex $d = \{A_0, A_1, A_2, A_3\} \in \Gamma^3$ be given. The conventional interpretation of the orientation will be that the middle two elements of an oriented 3-simplex are the elements of the common bond, the axis of rotation, and the order of atoms indicates the positive direction of the axis. The triangle with the first three elements is rotated into the triangle of the last three elements. So in the case of d^* defined as above, $\{A_1, A_2\}$ is the axis of rotation in the direction from $\{A_1\}$ to $\{A_2\}$, and t_0 is rotated into t_1. A signed angle, a **wedge angle**, between the half-plane containing t_0 and the half-plane containing t_1 is determined by $Z_{d^*}(R)$, which is defined as:

$$\begin{aligned} Z_{d^*}(R) = \quad & \mathbf{v}\cdot\mathbf{w} + i\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}, \text{ where } \mathbf{u} = \frac{\mathbf{R}_{A_2} - \mathbf{R}_{A_1}}{\|\mathbf{R}_{A_2} - \mathbf{R}_{A_1}\|}, \\ & \mathbf{v} = \frac{(1 - \mathbf{u}\mathbf{u}^T)(\mathbf{R}_{A_0} - \mathbf{R}_{A_1})}{\|(1 - \mathbf{u}\mathbf{u}^T)(\mathbf{R}_{A_0} - \mathbf{R}_{A_1})\|}, \text{ and } \mathbf{w} = \frac{(1 - \mathbf{u}\mathbf{u}^T)(\mathbf{R}_{A_3} - \mathbf{R}_{A_1})}{\|(1 - \mathbf{u}\mathbf{u}^T)(\mathbf{R}_{A_3} - \mathbf{R}_{A_1})\|}. \end{aligned}$$

It is easy to see that the vector $\mathbf{u}$ is well-defined. The vectors $\mathbf{v}$ and $\mathbf{w}$ are also well-defined which can be checked by a similar calculation as for $\mathbf{e}_2$ of a conformed pose. Recall that $[A_0, A_1, A_2, A_3] = [A_3, A_2, A_1, A_0]$ by the definition of orientation, but the definition of $Z_{d^*}(R)$, which should depend only on the equivalence class, seems to depend also on the permutation. $e = (A_0, A_1, A_2, A_3)$ and $e' = (A_3, A_2, A_1, A_0)$ are the only two permutations $(A_{\pi(0)}, A_{\pi(1)}, A_{\pi(2)}, A_{\pi(3)})$ of the same equivalence class such that $\{\{A_{\pi(0)}, A_{\pi(1)}, A_{\pi(2)}\}, \{A_{\pi(1)}, A_{\pi(2)}, A_{\pi(3)}\}\} = \{t_0, t_1\}$. So we need to check that $Z_e(R) = Z_{e'}(R)$, where $Z_e(R) = \mathbf{v}\cdot\mathbf{w}+i\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$ and $Z_{e'}(R) = \mathbf{v}'\cdot\mathbf{w}'+i\mathbf{u}'\cdot\mathbf{v}'\times\mathbf{w}'$. We can easily see that $\mathbf{u}' = -\mathbf{u}$, $\mathbf{v}' = \mathbf{w}$, and $\mathbf{w}' = \mathbf{v}$. Then $\mathbf{v}' \cdot \mathbf{w}' = \mathbf{w} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{w}$, and also $\mathbf{u}'\cdot\mathbf{v}'\times\mathbf{w}' = -\mathbf{u}\cdot\mathbf{w}\times\mathbf{v} = \mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$. Thus, we have $Z_e(R) = Z_{e'}(R)$. Indeed, since the two different permutations in the same equivalence class produce the same value for the function Z, $Z_{d^*}(R)$ is well-defined. So the wedge angle assigned does

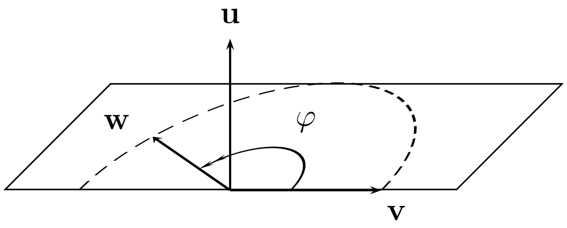


FIGURE 3. Positive Orientation of $\mathbf{u}, \mathbf{v}, \mathbf{w}$

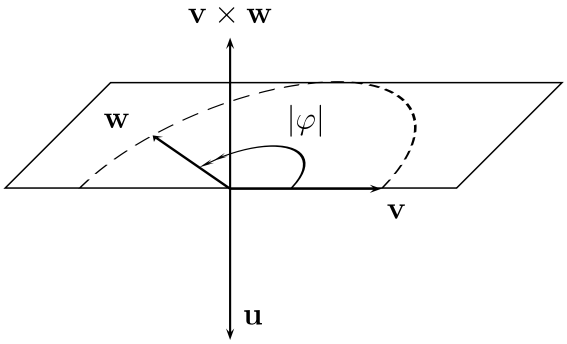


FIGURE 4. Negative Orientation of $\mathbf{u}, \mathbf{v}, \mathbf{w}$

not depend on whether t_0 is rotated into t_1, or t_1 is rotated into t_0 with a reversed axis of rotation.

The vector $\mathbf{u}$ is an unit vector on the common bond of the wedge corresponding to d in the direction of $\{A_1\}$ to $\{A_2\}$. Vectors $\mathbf{v}$ and $\mathbf{w}$ are the unit vectors perpendicular to $\mathbf{u}$ on the half-planes of triangles t_0 and t_1 respectively, and $\mathbf{v} \cdot \mathbf{w}$ gives the cosine of an angle between the half-planes; that is, $\mathbf{v} \cdot \mathbf{w} = \cos|\varphi|$, or $|\varphi| = \cos^{-1}(\mathbf{v} \cdot \mathbf{w})$ with $0 \leq |\varphi| \leq \pi$. $\mathbf{v}$ and $\mathbf{w}$ can be parallel, in which case $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = 0$ and $\varphi = 0$ or $\varphi = \pi$. If $\mathbf{v}$ and $\mathbf{w}$ are in the same direction, $\mathbf{v} \cdot \mathbf{w} = 1$, thus $\varphi = 0$. If they point to opposite from one another, then $\mathbf{v} \cdot \mathbf{w} = -1$, so $\varphi = \pi$. If $\mathbf{v}$ and $\mathbf{w}$ are not parallel, we have $0 < |\varphi| < \pi$. Since $\mathbf{v} \times \mathbf{w}$ gives a vector perpendicular to both $\mathbf{v}$ and $\mathbf{w}$, and $\mathbf{u}$ is perpendicular to both $\mathbf{v}$ and $\mathbf{w}$, $\mathbf{u}$ is either in the direction of $\mathbf{v} \times \mathbf{w}$

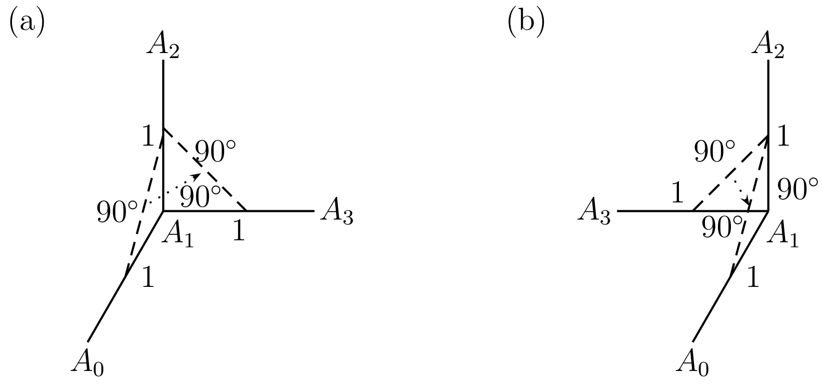


FIGURE 5. Two Orientations

or in the direction of $-(\mathbf{v} \times \mathbf{w})$. In the first case, where $\mathbf{v} \times \mathbf{w}$ and $\mathbf{u}$ are in the same direction, $\mathbf{v} \times \mathbf{w} = (\sin\varphi)\mathbf{u}$, hence $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = \sin\varphi > 0$, and so φ is positive (Figure 3). In the second case, we have $|\varphi| = \cos^{-1}(\mathbf{v} \cdot \mathbf{w})$ and $\mathbf{v} \times \mathbf{w} = (\sin|\varphi|)(-\mathbf{u})$, or $\mathbf{v} \times \mathbf{w} = (\sin(-|\varphi|))\mathbf{u}$. Then $\varphi = -|\varphi|$, and $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = \sin\varphi$ for $-\pi < \varphi < 0$ (Figure 4). So the first case and the second case are the mirror image of each other reflected about the plane determined by the vectors $\mathbf{v}$ and $\mathbf{w}$. Another way of writing $Z_{d^*}(R)$ is $Z_{d^*}(R) = \cos\varphi + i\sin\varphi = e^{i\varphi}$, giving a point on the unit circle in the set of complex numbers.

The orientations from different classes, $[A_0, A_1, A_2, A_3]$ and $[A_3, A_1, A_2, A_0]$ with the same bond lengths, angles, and wedge angles are said to have opposite **chirality** because one is the mirror image of the other and cannot be superimposed via rigid motions; that is, one orientation cannot be obtained by rotating and/or translating the other one. Suppose configuration R has the bond lengths and angles as shown in Figure 5 (a) with the orientation of the wedge $d^* = [A_0, A_1, A_2, A_3]$ so that $Z_{d^*}(R) = \frac{\pi}{2} \in (0, \pi)$. Just by changing the orientation to $\tilde{d}^* = [A_3, A_1, A_2, A_0]$, we get another configuration $\tilde{R}$ shown in Figure 5 (b) with $Z_{\tilde{d}^*}(\tilde{R}) = \frac{\pi}{2} \in (0, \pi)$. $\tilde{R}$ is the mirror image of R reflected about the plane containing the triangle $\{A_1, A_2, A_3]\}$, and $\tilde{R}$ cannot be obtained by rigid motions, which does not include the reflection about a plane. The label $e^{i\varphi}$ on $[A_0, A_1, A_2, A_3]$ has the same geometric content as the label $e^{-i\varphi}$ on $[A_3, A_1, A_2, A_0]$.

For $d^* = [A_0, A_1, A_2, A_3] \in \Gamma^3_*$ where $t_0 = \{A_0, A_1, A_2\}, t_1 = \{A_1, A_2, A_3\} \in \Gamma^2$, let $b_0 = \{A_1, A_2\}$, $b_0 \cup b_1 = t_0$, and $b_0 \cup b_2 = t_1$ for $b_0, b_1, b_2 \in \Gamma^2$. If $b_1 \cap b_2 = \emptyset$, then the tetrahedron is said to be a **dihedral**; otherwise, it is an **improper**. Another way of seeing this is if there is a path in (Γ^0, Γ^1) between $\{A_0\}$ and $\{A_3\}$ of length three, then it is a dihedral. If one atom is incident on all three bonds, then it is an improper. The **canonical orientation** of a dihedral is the one which follows that path in either way as $[A_0, A_1, A_2, A_3] = [A_3, A_2, A_1, A_0]$. An orientation of an improper determines and is determined by an ordering of the two triangles on which it is incident since the axis of rotation is always oriented from the atom $b_1 \cap b_2$ toward the other atom of b_0.

For a Z-system Γ^* and a configuration $R \in \mathcal{B}(\Gamma)$, we have defined $L.(R) \in (0, \infty)^{\Gamma^1}, C.(R) \in (-1, 1)^{\Gamma^2}$, and $Z.(R) \in (S^1)^{\Gamma^3_*}$. These three functions do not depend on which representative configuration R of a fixed conformation $G_a R$ one uses, as we will see in Section 1.5. We define the **parameter domain** $\mathcal{D}_P(\Gamma^*)$ to be:

$$\mathcal{D}_P(\Gamma^*) = (0, \infty)^{\Gamma^1} \times (-1, 1)^{\Gamma^2} \times (S^1)^{\Gamma^3_*}.$$

A typical element of $\mathcal{D}_P(\Gamma^*)$ is denoted by $\gamma = (L, C, Z)$. With γ specified, the edges of the trees of Γ^* are labeled with appropriate values; that is, the edges of the trees (Γ^0, Γ^1), (Γ^1, Γ^2), and (Γ^2, Γ^3) of Z-system Γ^* are labeled with the functions $L \in (0, \infty)^{\Gamma^1}$, $C \in (-1, 1)^{\Gamma^2}$ and $Z \in (S_1)^{\Gamma^3_*}$ respectively. The pair (Γ^*, γ) is called a **labeled Z-system**, and an element of $\mathcal{D}_P(\Gamma^*)$ is called a **labeling** of the Z-system. Suppose $\Gamma^* = (\Gamma^1, \Gamma^2, \Gamma^3_*)$ is Z-system and a site r is a root for Γ. Then define the mapping $\eta : \mathcal{B}(\Gamma) \to \mathcal{P} \times \mathcal{D}_P(\Gamma^*)$ by the rule:

$$R \mapsto (E_r(R), e \in \Gamma^1 \mapsto L_e(R), e \in \Gamma^2 \mapsto C_e(R), e^* \in \Gamma^3_* \mapsto Z_{e^*}(R)).$$

We will call this the **polyspherical trivialization** associated to (Γ^*, r). We also define the **coordinate domain** $\mathcal{D}_C(\Gamma)$ as:

$$\mathcal{D}_C(\Gamma) = G_a \backslash\backslash \mathcal{B}(\Gamma) = \{G_a R \mid R \in \mathcal{B}(\Gamma)\}.$$

Since the values of the functions L, C, and Z depend only on the specific conformation (see Section 1.5), we can also define the mapping $\hat{\eta} : \mathcal{D}_C(\Gamma) \to \mathcal{D}_P(\Gamma^*)$ by the rule:

$$\mathcal{O} \mapsto (e \in \Gamma^1 \mapsto \hat{L}_e(\mathcal{O}), e \in \Gamma^2 \mapsto \hat{C}_e(\mathcal{O}), e^* \in \Gamma^3_* \mapsto \hat{Z}_{e^*}(\mathcal{O}))$$

where $\hat{L}_e(\mathcal{O}) = L_e(R)$, $\hat{C}_e(\mathcal{O}) = C_e(R), \hat{Z}_{e^*}(\mathcal{O}) = Z_{e^*}(R)$ whenever $\mathcal{O} = G_a R$. We will call this the **polyspherical coordinate mapping** associated to the Z-system Γ^*. Note that $\hat{\eta}$ does not depend on the choice of a root r. Thus $\eta = (E_r, \hat{\eta} \circ \rho)$ where $\rho : \mathcal{B}(\Gamma) \to \mathcal{D}_C(\Gamma) : R \mapsto G_a R$.

Theorem. *Suppose $|\mathcal{N}| \geq 3$, $\Gamma^* = (\Gamma^1, \Gamma^2, \Gamma^3_*)$ is a Z-system on the set $\mathcal{N}$, r is a root for Γ, and $\eta : \mathcal{B}(\Gamma) \to \mathcal{P} \times \mathcal{D}_P(\Gamma^*)$ is the polyspherical trivialization. Then η is smooth one-to-one and onto with a smooth inverse mapping. Furthermore, the polyspherical coordinate mapping $\hat{\eta} : \mathcal{D}_C(\Gamma) \to \mathcal{D}_P(\Gamma^*)$ associated to Γ^* is also smooth one-to-one and onto, and its inverse is smooth.*

This theorem has been proven for the n-dimensional case in [8], where also the exact notion of smoothness, which involves the concept of manifold, is discussed.

So by the above theorem, we can conclude the following: given a set of mappings (L, C, Z) in $\mathcal{D}_P(\Gamma^*)$, the conformation $\mathcal{O}$ in $\mathcal{D}_C(\Gamma)$ is uniquely determined, and any conformation $\mathcal{O}$ in $\mathcal{D}_C(\Gamma)$ uniquely determines the mappings $(L, C, Z) \in \mathcal{D}_P(\Gamma^*)$.

1.5. Coordinate Transformation Matrix

Suppose Γ^* is a Z-system for which Γ is its underlying unoriented Z-system. Let vert $\mathcal{S}(\Gamma)$ denote the set of all sites r such that $s_k(r) \in \Gamma^k$ for $0 \leq k \leq 2$. These sites are said to be **associated with** or **from** the unoriented Z-system Γ. For $1 \leq k \leq 2$, let edge$_k\, \mathcal{S}(\Gamma)$ denote the set of all two-element subsets $\{(i_0, i_1, i_2), (j_0, j_1, j_2)\}$ of vert $\mathcal{S}(\Gamma)$ such that (j_0, j_1, j_2) is obtained from (i_0, i_1, i_2) by a transposition of the elements i_{k-1} and i_k. For $k = 3$, let edge$_k$ $\mathcal{S}(\Gamma)$ denote the set of all two-element subsets $\{(i_0, i_1, i), (i_0, i_1, i')\}$ of vert $\mathcal{S}(\Gamma)$ such that $\{i_0, i_1, i, i'\} \in \Gamma^3$. Let edge$_3\, \mathcal{S}(\Gamma^*)$ denote the set of all ordered pairs $((i_0, i_1, i), (i_0, i_1, i'))$ of distinct elements of vert $\mathcal{S}(\Gamma)$ such that $[i_0, i_1, i, i'] \in \Gamma^3_*$. Then define edge $\mathcal{S}(\Gamma) = \cup_{k=1}^{3}$edge$_k$ $\mathcal{S}(\Gamma)$ and edge $\mathcal{S}(\Gamma^*) = [\cup_{k=1}^{2}$edge$_k\, \mathcal{S}(\Gamma)] \cup$edge$_3\, \mathcal{S}(\Gamma^*)$. The traditional graph (vert $\mathcal{S}(\Gamma)$, edge $\mathcal{S}(\Gamma)$) is called the **undirected site graph** $\mathcal{S}(\Gamma)$, and it is connected whenever Γ is an unoriented Z-system [8]. (vert $\mathcal{S}(\Gamma)$, edge $\mathcal{S}(\Gamma^*)$) also forms a traditional graph, with elements of edge$_3\mathcal{S}(\Gamma^*)$ being directed, called the **site graph** $\mathcal{S}(\Gamma^*)$. See Section 1.8 for examples.

Define:

$$T_1(L) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ L & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \qquad L > 0;$$

$$T_2(C) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & C & \sqrt{1-C^2} & 0 \\ 0 & \sqrt{1-C^2} & -C & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad C \in (-1, 1);$$

$$T_3(Z) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & C & -S \\ 0 & 0 & S & C \end{pmatrix}, \qquad Z = C + iS \in S^1.$$

By definition, any edge $\{r, r'\}$ of the undirected site graph $\mathcal{S}(\Gamma)$ falls into one of three types, or sets, of edges: $\text{edge}_1\, \mathcal{S}(\Gamma)$, $\text{edge}_2\, \mathcal{S}(\Gamma)$, or $\text{edge}_3\, \mathcal{S}(\Gamma)$. Then, the following theorem, proven in [8], asserts that the matrix $\mathcal{A}_{r,r'}$ would be in the form of one of the three types of matrices above, meaning $E_{r'}(R) = E_r(R)\mathcal{A}_{r,r'}$ where the matrix $\mathcal{A}_{r,r'}$ is T_1, T_2, or T_3 with the appropriate argument.

Theorem. *Let Γ^* be a Z-system and $R \in \mathcal{B}(\Gamma)$. Suppose r and r' are 2-sites of Γ, and $\mathcal{A}_{r,r'}$ is the unique matrix in G_p such that $E_{r'}(R) = E_r(R)\mathcal{A}_{r,r'}$.*

(1) *If $\{r, r'\} \in$ $\text{edge}_1\, \mathcal{S}(\Gamma)$ and $e = s_1(r) = s_1(r')$, then $\mathcal{A}_{r,r'} = T_1(L_e(R))$.*

(2) *If $\{r, r'\} \in$ $\text{edge}_2\, \mathcal{S}(\Gamma)$ and $e = s_2(r) = s_2(r')$, then $\mathcal{A}_{r,r'} = T_2(C_e(R))$.*

(3) *If $r = (i_1, i_2, i_0)$ and $r' = (i_1, i_2, i_3)$ and $e^* = [i_0, i_1, i_2, i_3] \in \Gamma^3_*$, that is, $(r, r') \in$ $\text{edge}_3\, \mathcal{S}(\Gamma^*)$, then $\mathcal{A}_{r,r'} = T_3(Z_{e^*}(R))$.*

In a Z-system, the relation $E_{r'}(R) = E_r(R)\mathcal{A}_{r,r'}$ can be used to define the internal coordinates $L_e(R)$, $C_e(R)$, and $Z_{e^*}(R)$. First, suppose a bond $e = \{i_1, i_2\} \in \Gamma^1$ is given. Since $e \in \Gamma^1$ is a vertex of the connected graph (Γ^1, Γ^2), we know there exists a triangle $t = \{i_0, i_1, i_2\} \in \Gamma^2$ on which e is incident. So we are able to obtain two sites, $r = (i_1, i_2, i_0)$ and $r' = (i_2, i_1, i_0)$. Then we know the unique matrix $\mathcal{A}_{r,r'}$ such that $E_r'(R) = E_r(R)\mathcal{A}_{r,r'}$. So by the above theorem, we have $T_1(L_e(R)) = \mathcal{A}_{r,r'}$, and $L_e(R)$ is uniquely determined. Next, if we are given a triangle $e = \{i_0, i_1, i_2\} \in \Gamma^2$, then we have two bonds $b = \{i_0, i_1\}$ and $b' = \{i_1, i_2\}$ both in Γ^1 that are incident on e. So we have the two sites $r = (i_1, i_0, i_2)$ and $r' = (i_1, i_2, i_0)$ such that $\{r, r'\} \in \text{edge}_2\mathcal{S}(\Gamma)$ and $s_2(r) = s_2(r') = e$, and we can find $\mathcal{A}_{r,r'}$ such that $E_r'(R) = E_r(R)\mathcal{A}_{r,r'}$. Then again by the theorem, we conclude that $T_2(C_e(R)) = \mathcal{A}_{r,r'}$, determining $C_e(R)$ uniquely.

Now suppose we are given an orientation of a tetrahedron and two triangles that are incident on that tetrahedron. So without loss of generality, let $e^* = [i_0, i_1, i_2, i_3] \in \Gamma^3_*$ and $t = \{i_0, i_1, i_2\}, t' = \{i_1, i_2, i_3\} \in \Gamma^2$. We can easily obtain two sites by using these two triangles. The common bond of t and t' is the bond $\{i_1, i_2\}$ with the direction of the axis of orientation from i_1 to i_2. So the two sites would be $r = (i_1, i_2, i_0)$ and $r' = (i_1, i_2, i_3)$ and $(r, r') \in \text{edge}_3\mathcal{S}(\Gamma^*)$. Then the matrix $\mathcal{A}_{r,r'}$ such that $E'_r(R) = E_r(R)\mathcal{A}_{r,r'}$ exists and is unique, and by the theorem, $T_3(Z_{e^*}(R)) = \mathcal{A}_{r,r'}$, which uniquely determines $Z_{e^*}(R)$.

From the following, we will now see that $\mathcal{A}_{r,r'}$ does not depend on the poses given by the particular configuration R. Since $E_{r'}(R) = E_r(R)\mathcal{A}_{r,r'}(R)$, we let $(\mathbf{b}, A) \in G_a$ act on the both sides from the left; that is, $(\mathbf{b}, A) \cdot E_{r'}(R) = (\mathbf{b}, A) \cdot [E_r(R)\mathcal{A}_{r,r'}(R)]$. Then using the characteristics of the pose mentioned in Section 1.2,

$$\begin{aligned} E_{r'}((\mathbf{b}, A)R) &= (\mathbf{b}, A) \cdot E_{r'}(R) \\ &= (\mathbf{b}, A) \cdot [E_r(R)\mathcal{A}_{r,r'}(R)] \\ &= [(\mathbf{b}, A) \cdot E_r(R)]\mathcal{A}_{r,r'}(R) \\ &= E_r((\mathbf{b}, A)R)\mathcal{A}_{r,r'}(R). \end{aligned}$$

So $E_{r'}((\mathbf{b}, A)R) = E_r((\mathbf{b}, A)R)\mathcal{A}_{r,r'}(R)$. Because of the uniqueness of such matrix $\mathcal{A}_{r,r'}(R)$, we have $\mathcal{A}_{r,r'}((\mathbf{b}, A)R) = \mathcal{A}_{r,r'}(R)$. Therefore, we conclude that the matrix $\mathcal{A}_{r,r'}$ does not depend on configurations except through the associated conformation.

This implies that the T-matrices are independent of the representative of configuration of a fixed conformation since the $\mathcal{A}_{r,r'}$ is a T-matrix. Then since each T-matrix depends only on one argument, we conclude that the arguments of T-matrices, the functions L, C, Z of the parameter domain, are independent of the representative of configuration of a fixed conformation.

1.6. Tethering and Gluing

Let Γ and Λ be unoriented Z-systems of two reactants such that $\Gamma^0 \cap \Lambda^0 = \emptyset$. As the two reactants approach each other in space to react, it becomes necessary to form a new Z-system $\Gamma \oplus_\mu \Lambda$, using information μ. This operation is called **tethering**. Let $\mu = \{(i_0, i_1, i_2), (j_0, j_1, j_2)\}$ where (i_0, i_1, i_2) is a site of Γ and (j_0, j_1, j_2) is of Λ. In this case, $\{i_0, j_0\}$ is a new bond being formed between the two reactants. Figure 6 shows the two 2-sites (i_0, i_1, i_2) and (j_0, j_1, j_2) being tethered where the solid lines represent the bonds. The thickest black line shows the new bond, and the two gray lines in each site show two possible bonds that could be in the Z-system, but only one of these two bonds is actually in the Z-system. We define the new unoriented Z-system as follows:

$$
\begin{aligned}
(\Gamma \oplus_\mu \Lambda)^0 &= \Gamma^0 \cup \Lambda^0 \\
(\Gamma \oplus_\mu \Lambda)^1 &= \Gamma^1 \cup \Lambda^1 \cup \{\{i_0, j_0\}\} \\
(\Gamma \oplus_\mu \Lambda)^2 &= \Gamma^2 \cup \Lambda^2 \cup \{\{i_0, i_1, j_0\}, \{i_0, j_0, j_1\}\} \\
(\Gamma \oplus_\mu \Lambda)^3 &= \Gamma^3 \cup \Lambda^3 \cup \{\{i_0, i_1, i_2, j_0\}, \{i_0, i_1, j_0, j_1\}, \{i_0, j_0, j_1, j_2\}\}.
\end{aligned}
$$

With these definitions, $((\Gamma \oplus_\mu \Lambda)^0, (\Gamma \oplus_\mu \Lambda)^1), ((\Gamma \oplus_\mu \Lambda)^1, (\Gamma \oplus_\mu \Lambda)^2)$, and $((\Gamma \oplus_\mu \Lambda)^2, (\Gamma \oplus_\mu \Lambda)^3)$ are all trees [8], so that $\Gamma \oplus_\mu \Lambda$ becomes an unoriented Z-system. If Γ^* and Λ^* are oriented Z-systems, then $\Gamma \oplus_\mu \Lambda$ can be made into an oriented Z-system, denoted by $\Gamma^* \oplus_\mu \Lambda^*$, in a natural way. The central tetrahedron is a dihedral so we assign it the canonical orientation $[i_1, i_0, j_0, j_1]$. For the other two, we assign the orientations $[j_0, i_0, i_1, i_2]$ and $[i_0, j_0, j_1, j_2]$ so that the orientation is canonical if the tetrahedron is dihedral, and if the tetrahedron is improper, the new triangle is to be

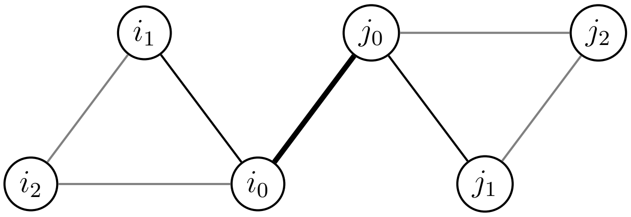


Figure 6. Tethering a 2-Site to a 2-Site

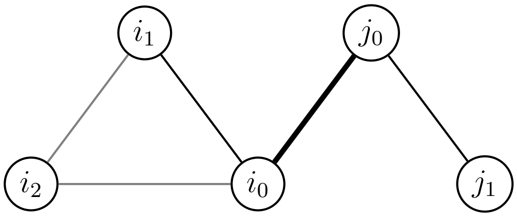


FIGURE 7. Tethering a 1-Site to a 2-Site

rotated into the old triangle. So $(\Gamma^* \oplus_\mu \Lambda^*)^3_*$ is defined as:

$$(\Gamma^* \oplus_\mu \Lambda^*)^3_* = \Gamma^3_* \cup \Lambda^3_* \cup \{[j_0, i_0, i_1, i_2], [i_1, i_0, j_0, j_1], [i_0, j_0, j_1, j_2]\}.$$

For labeled Z-systems, we need to specify the numerical labels of the six new simplices to determine the relative position and the orientation of the reactants. If (Γ^*, γ) and (Λ^*, λ) are labeled Z-systems, then a labeled Z-system $(\Gamma^* \oplus_\mu \Lambda^*, \delta)$ can be defined, where δ agrees with γ on Γ^*, λ on Λ^*, and also gives the assignments of the numerical labels for the new simplices.

It is also possible to tether a diatomic or monatomic component to a multi-atomic component of a GZ-system Γ on $\mathcal{N}$ to obtain a larger multi-atomic component in a new GZ-system Λ on $\mathcal{N}$. Let $\mathcal{C} \subset \mathcal{N}$ be a multi-atomic component with its site (i_0, i_1, i_2), and let $\mathcal{D} \subset \mathcal{N}$ be a diatomic component with the site (j_0, j_1). Then the tethered unoriented Z-system $\Lambda_{\mathcal{C}\cup\mathcal{D}} = \Gamma_\mathcal{C} \oplus_\mu \Gamma_\mathcal{D}$ with $\mu = \{(i_0, i_1, i_2), (j_0, j_1)\}$ is defined as:

$$\begin{aligned}
(\Gamma_\mathcal{C} \oplus_\mu \Gamma_\mathcal{D})^0 &= \Gamma^0_\mathcal{C} \cup \Gamma^0_\mathcal{D} \\
(\Gamma_\mathcal{C} \oplus_\mu \Gamma_\mathcal{D})^1 &= \Gamma^1_\mathcal{C} \cup \Gamma^1_\mathcal{D} \cup \{\{i_0, j_0\}\} \\
(\Gamma_\mathcal{C} \oplus_\mu \Gamma_\mathcal{D})^2 &= \Gamma^2_\mathcal{C} \cup \Gamma^2_\mathcal{D} \cup \{\{i_0, i_1, j_0\}, \{i_0, j_0, j_1\}\} \\
(\Gamma_\mathcal{C} \oplus_\mu \Gamma_\mathcal{D})^3 &= \Gamma^3_\mathcal{C} \cup \Gamma^3_\mathcal{D} \cup \{\{i_0, i_1, i_2, j_0\}, \{i_0, i_1, j_0, j_1\}\}.
\end{aligned}$$

For the oriented Z-system, we assign the orientation in the same way as before, so

$$(\Gamma^*_\mathcal{C} \oplus_\mu \Gamma^*_\mathcal{D})^3_* = (\Gamma_\mathcal{C})^3_* \cup (\Gamma_\mathcal{D})^3_* \cup \{[j_0, i_0, i_1, i_2], [i_1, i_0, j_0, j_1]\}.$$

Figure 7 shows 1-site (j_0, j_1) being tethered to a 2-site (i_0, i_1, i_2).

Similarly, if $\{j_0\} = \mathcal{M} \subset \mathcal{N}$ is monatomic with the site (j_0), we define the tethered Z-system with $\mu = \{(i_0, i_1, i_2), (j_0)\}$ as:

$$\begin{aligned}
(\Gamma_{\mathcal{C}} \oplus_\mu \Gamma_{\mathcal{M}})^0 &= \Gamma^0_{\mathcal{C}} \cup \Gamma^0_{\mathcal{M}} \\
(\Gamma_{\mathcal{C}} \oplus_\mu \Gamma_{\mathcal{M}})^1 &= \Gamma^1_{\mathcal{C}} \cup \Gamma^1_{\mathcal{M}} \cup \{\{i_0, j_0\}\} \\
(\Gamma_{\mathcal{C}} \oplus_\mu \Gamma_{\mathcal{M}})^2 &= \Gamma^2_{\mathcal{C}} \cup \Gamma^2_{\mathcal{M}} \cup \{\{i_0, i_1, j_0\}\} \\
(\Gamma_{\mathcal{C}} \oplus_\mu \Gamma_{\mathcal{M}})^3 &= \Gamma^3_{\mathcal{C}} \cup \Gamma^3_{\mathcal{M}} \cup \{\{i_0, i_1, i_2, j_0\}\} \\
(\Gamma^*_{\mathcal{C}} \oplus_\mu \Gamma^*_{\mathcal{M}})^3_* &= (\Gamma_{\mathcal{C}})^3_* \cup (\Gamma_{\mathcal{M}})^3_* \cup \{[j_0, i_0, i_1, i_2]\}.
\end{aligned}$$

As before we need to specify the numerical values for the new simplices to obtain a labeled tethered Z-system.

In the simulation of a chemical reaction, tethering is useful especially when the molecules are very close to each other and are ready to react (form a new chemical bond) and when the new products are about to come apart. It is not good choice to use tethering to describe molecules that are just floating around randomly because one of the new bond angles might become 0° or 180°. In such a case, some wedge angles might become undefined; or the bond angle might become close to 0° or 180° in which case some wedge angles may vary wildly, leading to numerical instability. For a system of molecules in which there is a possibility of bond angles becoming 0° or 180°, we employ separate components of a GZ-system. Tethering is useful also in the docking of two parts such as two helices. For examples of tethering, see Section 1.8.

When building a model for a large molecule, it is often desirable to build it from smaller pieces by **gluing** them together even though the process does not conserve atoms and hence is not chemically accurate. Suppose Γ^* and Λ^* are Z-systems on $\mathcal{N}$ and $\mathcal{M}$, the two pieces being glued, where $\Gamma^0 \cap \Lambda^0 = \emptyset$. To start, we choose sites (i_0, i_1, i_2) from Γ and (j_0, j_1, j_2) from Λ. We require that $\{i_0\}$ is a leaf in (Γ^0, Γ^1) and $\{j_0\}$ is a leaf in (Λ^0, Λ^1). Define $\mathcal{N} *_\mu \mathcal{M} = (\mathcal{N} \backslash \{i_0\}) \cup (\mathcal{M} \backslash \{j_0\})$ with $\mu =$

$\{(i_0, i_1, i_2), (j_0, j_1, j_2)\}$. The atoms $\{i_0\}$ and $\{j_0\}$ are destroyed by gluing. A new Z-system $\Gamma *_\mu \Lambda$ on $\mathcal{N} *_\mu \mathcal{M}$ is obtained using mappings $\iota_\mathcal{N} : \mathcal{N} \to \mathcal{N} *_\mu \mathcal{M}$ and $\iota_\mathcal{M} : \mathcal{M} \to \mathcal{N} *_\mu \mathcal{M}$ defined as:

$$\iota_\mathcal{N}(i) = \begin{cases} i & i \in \mathcal{N}\backslash\{i_0\} \\ j_1 & i = i_0 \end{cases}, \qquad \iota_\mathcal{M}(j) = \begin{cases} j & j \in \mathcal{M}\backslash\{j_0\} \\ i_1 & j = j_0 \end{cases}.$$

The mappings $\iota_\mathcal{N}$ and $\iota_\mathcal{M}$ can be applied to any subset of their domains by applying the mappings to each element of the subset. For example, $\iota_\mathcal{N}(\{i_{l_0}, i_{l_1}, \ldots, i_{l_k}\}) = \{\iota_\mathcal{N}(i_{l_0}), \iota_\mathcal{N}(i_{l_1}), \ldots, \iota_\mathcal{N}(i_{l_k})\}$. Also those mappings map ordered tuples and the oriented 3-simplices in an obvious way. So we define the new Z-system as the following:

$$\begin{aligned} (\Gamma *_\mu \Lambda)^0 &= \{\iota_\mathcal{N}(a) \mid a \in \Gamma^0\} \cup \{\iota_\mathcal{M}(a) \mid a \in \Lambda^0\} = \binom{\mathcal{N} *_\mu \mathcal{M}}{1} \\ (\Gamma *_\mu \Lambda)^1 &= \{\iota_\mathcal{N}(b) \mid b \in \Gamma^1\} \cup \{\iota_\mathcal{M}(b) \mid b \in \Lambda^1\} \\ (\Gamma *_\mu \Lambda)^2 &= \{\iota_\mathcal{N}(t) \mid t \in \Gamma^2\} \cup \{\iota_\mathcal{M}(t) \mid t \in \Lambda^2\}. \end{aligned}$$

For the resulting piece to be a Z-system, a new tetrahedron $\{i_2, i_1, j_1, j_2\}$ should be added, and since this is a dihedral, we assign a canonical orientation. So:

$$\begin{aligned} (\Gamma *_\mu \Lambda)^3 &= \{\iota_\mathcal{N}(d) \mid d \in \Gamma^3\} \cup \{\iota_\mathcal{M}(d) \mid d \in \Lambda^3\} \cup \{\{i_2, i_1, j_1, j_2\}\} \\ (\Gamma^* *_\mu \Lambda^*)^3_* &= \{\iota_\mathcal{N}(d^*) \mid d^* \in \Gamma^3_*\} \cup \{\iota_\mathcal{M}(d^*) \mid d^* \in \Lambda^3_*\} \cup \{[i_2, i_1, j_1, j_2]\}. \end{aligned}$$

This determines a well-defined Z-system [8]. If (Γ^*, γ) for $\gamma = (L^\gamma, C^\gamma, Z^\gamma)$ and (Λ^*, λ) for $\lambda = (L^\lambda, C^\lambda, Z^\lambda)$ are labeled Z-systems, then $(\Gamma^* *_\mu \Lambda^*, \delta)$ can be defined where $\delta = (L^\delta, C^\delta, Z^\delta)$ satisfies the following:

$$L^\delta(\tilde{b}) = \begin{cases} L^\gamma(b) & \text{if } \tilde{b} = \iota_\mathcal{N}(b) \text{ for } b \in \Gamma^1\backslash\{\{i_0, i_1\}\} \\ L^\lambda(b) & \text{if } \tilde{b} = \iota_\mathcal{M}(b) \text{ for } b \in \Lambda^1\backslash\{\{j_0, j_1\}\} \end{cases},$$

$$C^\delta(\tilde{t}) = \begin{cases} C^\gamma(t) & \text{if } \tilde{t} = \iota_\mathcal{N}(t) \text{ for } t \in \Gamma^2 \\ C^\lambda(t) & \text{if } \tilde{t} = \iota_\mathcal{M}(t) \text{ for } t \in \Lambda^2 \end{cases},$$

$$Z^{\delta}(\tilde{d^*}) \quad = \quad \begin{cases} Z^{\gamma}(d^*) & \text{if } \tilde{d^*} = \iota_{\mathcal{N}}(d^*) \text{ for } d^* \in \Gamma^3_* \\ Z^{\lambda}(d^*) & \text{if } \tilde{d^*} = \iota_{\mathcal{M}}(d^*) \text{ for } d^* \in \Lambda^3_* \end{cases} .$$

To complete the labeled Z-system, a wedge angle for the new tetrahedron and the bond length of $\{i_1, j_1\}$ should be assigned. For examples of gluing, see Section 4.1.

1.7. Conformations and Labeled GZ-systems

The undirected site graph $\mathcal{S}(\Gamma)$ for a unoriented Z-system Γ discussed in Section 1.5 is connected. Thus we can find a path in $\mathcal{S}(\Gamma)$ between any two distinct sites of Γ. Let r and r' be 2-sites from the Z-system Γ and p be the path from r to r' of the undirected site graph $\mathcal{S}(\Gamma)$. Then as previously discussed in Section 1.2, if $R \in \mathcal{B}(\Gamma)$, there exists a unique matrix $\mathcal{A}_{r,r'}$ in G_p such that $E'_r(R) = E_r(R)\mathcal{A}_{r,r'}$. Suppose $r_0, r_1, \ldots, r_m$ is the sequence of vertices in the path p where $r_0 = r$ and $r_m = r'$. Then there is also a unique matrix $\mathcal{A}_{r_{j-1},r_j}$ such that $E_{r_j}(R) = E_{r_{j-1}}(R)\mathcal{A}_{r_{j-1},r_j}$ in G_p for each $j = 1, \ldots, m$. Then we have a sequence of matrices with which we relate the two arbitrary chosen sites $r = r_0$ and $r' = r_m$ as:

$$E_{r'}(R) = E_r(R)\mathcal{A}_{r_0,r_1}\mathcal{A}_{r_1,r_2}\ldots\mathcal{A}_{r_{m-1},r_m}.$$

So the unique matrix $\mathcal{A}_{r,r'}$ such that $E_{r'}(R) = E_r(R)\mathcal{A}_{r,r'}$ is:

$$\mathcal{A}_{r,r'} = \mathcal{A}_{r_0,r_1}\mathcal{A}_{r_1,r_2}\ldots\mathcal{A}_{r_{m-1},r_m}$$

for $r = r_0$ and $r' = r_m$. In this way, given a 2-site with its pose and the labeling of the Z-system Γ^*, we are able find the pose of any other site on Γ^* since by the theorem of Section 1.5, each $\mathcal{A}_{r_{j-1},r_j}$ will be a T-matrix whose argument is taken from the labeling.

A 2-site in the context of biomolecular system is a set consisting of a particular atom and a bond incident on that atom and a triangle of which the bond is a part. That is, a site (flag) $r = \{a, b, t\}$, or $r = (A_0, A_1, A_2)$ as an ordered triple, is such that $a = \{A_0\} \in \Gamma^0$, $b = \{A_0, A_1\} \in \Gamma^1$, and $t = \{A_0, A_1, A_2\} \in \Gamma^2$. Since $R \in \mathcal{B}(\Gamma)$ implies $\{\mathbf{R}_{A_0}, \mathbf{R}_{A_1}, \mathbf{R}_{A_2}\}$ is noncollinear, we have a well-defined pose associated to the site. A 1-site would be $\{a, b\}$, or (A_0, A_1) as an ordered pair, for $a = \{A_0\} \in \Gamma^0$ and $b = \{A_0, A_1\} \in \Gamma^1$; and a 0-site is $\{a\}$, or (A_0) where $A_0 \in \Gamma^0$.

We say that a k-site r is **associated** to a component $\mathcal{C}$ of GZ-system Γ if $s_0(r) \in \Gamma^0_{\mathcal{C}}, s_1(r) \in \Gamma^1_{\mathcal{C}}, \ldots, s_k(r) \in \Gamma^k_{\mathcal{C}}$ and $k+1 = \min\{3, |\mathcal{C}|\}$. For example, a site associated to a multi-atomic component would be a 2-site, and a site associated to a diatomic component is a 1-site. Let Γ^* be an oriented GZ-system and Γ be its underlying unoriented GZ-system, with at least two components. Let Comp be the set of all components in Γ^*. If r is a 2-site associated to a multi-atomic component $\mathcal{C}$ and r' is a k-site associated to another (distinct) component $\mathcal{C}'$ for $k = 0, 1, 2$, then an ordered pair (r, r') is called a **linkage** from the 2-site r to the k-site r'. In this situation, we say (r, r') is **incident** on $\mathcal{C}$ and $\mathcal{C}'$. Let Link be a collection of linkages with the following two properties:

(1) If $\mathcal{C}$ and $\mathcal{C}'$ are distinct components in Comp, then there is at most one linkage $(r, r') \in \text{Link}$ incident on $\mathcal{C}$ and $\mathcal{C}'$.

(2) $(\text{Comp}, \text{Link})$, understood as an abstract graph with the above incidence relation, is a tree (called the **component/linkage tree**).

It follows that every diatomic or monatomic component will be a leaf of the tree $(\text{Comp}, \text{Link})$. For each linkage $(r, r') \in \text{Link}$ of the component/linkage tree, we will attempt to label it with a matrix $\mathcal{A}_{r,r'} \in G_p^{(k)}$ where r' is a k-site. If $R \in \mathcal{B}(\Gamma)$ and r and r' are sites of Γ, then $\mathcal{A}_{r,r'}(R)$ is uniquely determined so that $E_{r'}(R) = E_r(R)\mathcal{A}_{r,r'}(R)$. In fact, $\mathcal{A}_{r,r'}((\mathbf{b}, A)R) = \mathcal{A}_{r,r'}(R)$, so $\mathcal{A}_{r,r'}$ only depends on the conformation $G_a R \in \mathcal{D}_C(\Gamma) = G_a \backslash\!\backslash \mathcal{B}(\Gamma)$. One method of finding such an $\mathcal{A}$-matrix without first giving a configuration R is demonstrated in the next section.

Having a concept of a function $\mathcal{A}$ which assigns a matrix to each linkage, we have the following theorem for describing the conformation of a GZ-system.

Theorem. *Suppose* $|\mathcal{N}| \geq 3$ *and* (Γ^*, r) *is a rooted GZ-system on* $\mathcal{N}$ *and suppose* $(\text{Comp}, \text{Link})$ *is a component/linkage tree for* Γ^**. Define*

$$\mathcal{D}_P(\Gamma^*) = (0, \infty)^{\Gamma^1} \times (-1, 1)^{\Gamma^2} \times (S^1)^{\Gamma^3_*}$$

and

$$\mathcal{A}_P(\mathrm{Link}) = \prod_{\substack{(r,r')\in\mathrm{Link}\\ r' \textit{ is a k-site}}} G_p^{(k)}$$

Then the mapping

$$\begin{aligned} \eta: \ \mathcal{B}(\Gamma) \to \quad & \mathcal{P} \times \mathcal{D}_P(\Gamma^*) \times \mathcal{A}_P(\mathrm{Link}) : \\ R \mapsto \qquad & (E_r(R), e \in \Gamma^1 \mapsto L_e(R), e \in \Gamma^2 \mapsto C_e(R), e^* \in \Gamma^3_* \mapsto Z_{e^*}(R), \\ & \quad (r, r') \in \mathrm{Link} \mapsto \mathcal{A}_{r,r'}(R)) \end{aligned}$$

is smooth one-to-one and onto and its inverse is smooth. Also

$$\begin{aligned} \hat{\eta}: \ \mathcal{D}_C(\Gamma) \to \quad & \mathcal{D}_P(\Gamma^*) \times \mathcal{A}_P(\mathrm{Link}) : \\ \mathcal{O} \mapsto \qquad & (e \in \Gamma^1 \mapsto \hat{L}_e(\mathcal{O}), e \in \Gamma^2 \mapsto \hat{C}_e(\mathcal{O}), e^* \in \Gamma^3_* \mapsto \hat{Z}_{e^*}(\mathcal{O}), \\ & \quad (r, r') \in \mathrm{Link} \mapsto \hat{\mathcal{A}}_{r,r'}(\mathcal{O})) \end{aligned}$$

is smooth one-to-one and onto with a smooth inverse. $\hat{\eta}$ *is independent of the root* r.

Proof. (Sketch) The natural way to construct the inverse of η is to start from the root component. Let $(E_r, \gamma, \mathcal{A}) \in \mathcal{P} \times \mathcal{D}_P(\Gamma^*) \times \mathcal{A}_P(\mathrm{Link})$ be given. We must construct $R \in \mathcal{B}(\Gamma))$ such that $\eta(R) = (E_r, \gamma, \mathcal{A})$. Since (Γ^*, r) is a rooted GZ-system, we consider the root r and the component $\mathcal{C}$ to which r is associated. $\mathcal{C}$ is necessarily multiatomic. Then $(\Gamma^*_{\mathcal{C}}, r)$ is a rooted Z-system, and $\gamma_{\mathcal{C}}$ is its labeling where $\gamma_{\mathcal{C}}$ is the labeling γ restricted to $\Gamma^*_{\mathcal{C}}$. We apply the theorem for the Z-system in Section 1.4. So given $(E_r, \Gamma_{\mathcal{C}}) \in \mathcal{P} \times \mathcal{D}_P(\Gamma^*_{\mathcal{C}})$, we apply the theorem of Section 1.4 to obtain $R_{\mathcal{C}} \in \mathcal{B}(\Gamma^*_{\mathcal{C}})$ such that $\eta(R_{\mathcal{C}}) = (E_r, \gamma_{\mathcal{C}})$. From $R_{\mathcal{C}}$ and any site r_1 associated to the component $\mathcal{C}$, the pose $E_{r_1}(R_{\mathcal{C}})$ is determined. So for the linkage $(r_1, r'_1) \in \mathrm{Link}$ where r_1 is a site on $\mathcal{C}$ and r'_1 is a site on $\mathcal{C}' \neq \mathcal{C}$, we know the pose $E_{r_1}(R_{\mathcal{C}})$ at r_1 by the theorem. Then by applying the matrix $\mathcal{A}_{r_1,r'_1}$ assigned to the linkage, the pose $E_{r'_1} = E_{r_1}(R_{\mathcal{C}})\mathcal{A}_{r_1,r'_1}$ is obtained. Now considering the component $\mathcal{C}'$ and the site r'_1, $(\Gamma^*_{\mathcal{C}'}, r'_1)$ is again a rooted Z-system (provided that $\mathcal{C}'$ i multiatomic) with the labeling $\gamma_{\mathcal{C}'}$. So by applying the theorem of Section 1.4 to this Z-system, we obtain the configuration $R_{\mathcal{C}'}$ for this component. Then, we are able to find the pose $E_{r'_2}(R_{\mathcal{C}'})$

for the linkage $(r_2', r_2'') \in \mathrm{Link}$ where r_2' is a site associated to $\mathcal{C}'$ and r_2'' to $\mathcal{C}''$. If $\mathcal{C}'$ is diatomic or monatomic, then $E_{r_1'}$ already defines an extension of R to include the atoms of $\mathcal{C}'$. So by repeating this process till all the components are covered, we obtain the configuration for the whole GZ-system. One problem is if there were more than one way to reach a component from the root component. However, since we require $(\mathrm{Comp}, \mathrm{Link})$ to be a tree, no cycle is possible; thus, each site in any linkage in Link is reached exactly one way. Also since $(\mathrm{Comp}, \mathrm{Link})$ is also connected, no component is left with its configuration undetermined. $\square$

By this theorem, given a set of mappings $(L, C, Z, \mathcal{A})$ where $(L, C, Z) \in \mathcal{D}_P(\Gamma^*)$ and $\mathcal{A} \in \mathcal{A}_P(\mathrm{Link})$ for a GZ-system Γ^*, the conformation $\mathcal{O}$ in $\mathcal{D}_C(\Gamma)$ is uniquely determined. Also, given a conformation $\mathcal{O}$ in $\mathcal{D}_C(\Gamma)$, we can determine the unique mappings $(L, C, Z, \mathcal{A})$ such that $(L, C, Z) \in \mathcal{D}_P(\Gamma^*)$ and $\mathcal{A} \in \mathcal{A}_P(\mathrm{Link})$. For this reason when we speak of a GZ-system, we mean not only Γ^* but also a choice of the component/linkage graph $(\mathrm{Comp}, \mathrm{Link})$; and a labeled GZ-system will include a specification of $(L, C, Z, \mathcal{A})$.

1.8. Example of a GZ-System

To understand some of the concepts discussed in this chapter, especially how to find the coordinate transformation matrix to assign to a given linkage without knowing a configuration of the entire system, we will take a concrete example, a system of sodium bicarbonate being dissolved in water. Assume that we are interested in a certain state of that system where a water molecule is coordinating a sodium ion and a hydroxide ion and a carbonic acid molecule by hydrogen bonds. The GZ-system of the system in that state is depicted in Figure 2. The set of atoms that are incident on the two bonds of an angle is the triangle in the GZ-system. Since exactly two bonds of the same triangle can be in the GZ-system, exactly one angle is associated with that triangle. So the triangles and the angles are one-to-one correspondence. A dotted line is for the wedge (tetrahedron) with its label given in degrees in a dotted line box. In a GZ-system, two triangles can form a tetrahedron only when those triangles have two of their elements in common. This is because a wedge exists between two angles only when they share a bond. The tetrahedra and the wedges are also one-to-one correspondence [8]. The dihedrals $[O_5, C, O_3, H_3]$ and $[O_5, C, O_4, H_4]$ are assumed to have the canonical orientation in this figure. The improper $[O_3, C, O_5, O_4]$ is shown as a dotted arrow, and the direction of the arrow indicates which triangle is rotated into which, and this information determines the orientation of the tetrahedron as discussed in Section 1.4.

For the system of molecules in Figure 2, we have a GZ-system Γ^* on $\mathcal{N}$, the set of all atom names that are in the system, so

$$\mathcal{N} = \{Na, O_1, H_1, H', O_2, H_2, C, O_3, H_3, O_4, H_4, O_5\}.$$

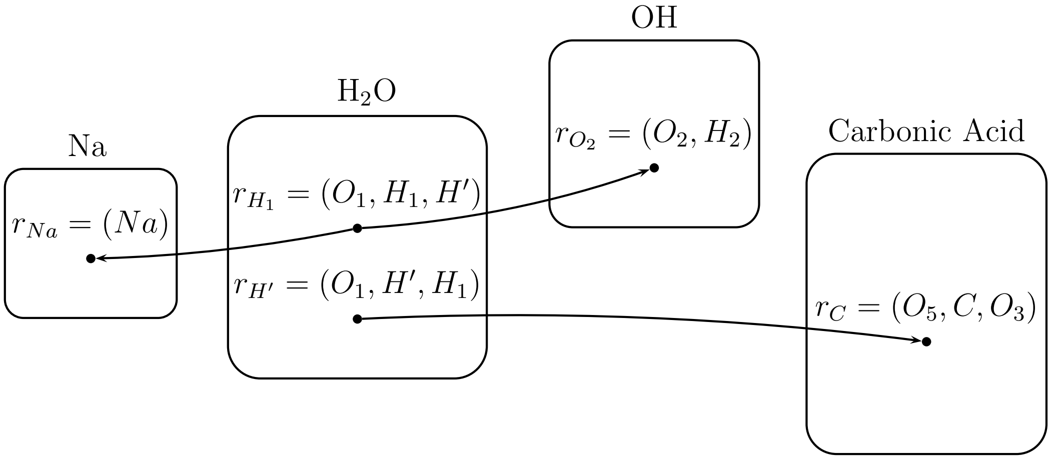


FIGURE 8. Component/Linkage Tree for the GZ-system

This GZ-system consists of different types of components: a monatomic component Na= $\{Na\}$, a diatomic component OH= $\{O_2, H_2\}$, and two multi-atomic components H_2O= $\{O_1, H_1, H'\}$ and Carbonic Acid= $\{C, O_3, H_3, O_4, H_4, O_5\}$.

The unoriented GZ-system is defined as follows:

$$\begin{aligned}
\Gamma^0 &= \{\{Na\}, \{O_1\}, \{H_1\}, \{H'\}, \{O_2\}, \{H_2\}, \{C\}, \{O_3\}, \{H_3\}, \{O_4\}, \{H_4\}, \{O_5\}\} \\
\Gamma^1 &= \{\{O_1, H_1\}, \{O_1, H'\}, \{O_2, H_2\}, \{C, O_3\}, \{O_3, H_3\}, \{C, O_4\}, \{O_4, H_4\}, \{C, O_5\}\} \\
\Gamma^2 &= \{\{O_1, H_1, H'\}, \{C, O_3, H_3\}, \{C, O_4, H_4\}, \{C, O_3, O_5\}, \{C, O_4, O_5\}\} \\
\Gamma^3 &= \{\{O_3, C, O_5, O_4\}, \{H_3, O_3, C, O_5\}, \{H_4, O_4, C, O_5\}\},
\end{aligned}$$

and for the oriented GZ-system,

$$\Gamma^3_* = \{[O_3, C, O_5, O_4], [H_3, O_3, C, O_5], [H_4, O_4, C, O_5]\}.$$

The labels for the Carbonic Acid are taken from the values for a formic acid in [18], and other labels are rough estimates of what they could be. Those properties vary depending on the environment in which the molecules are situated. The distances between the atoms of the hydrogen bonds are also estimated based on the information available for one type of hydrogen bonds between oxygen and hydroxide [22].

The component/linkage tree, $(\mathrm{Comp}, \mathrm{Link})$ could be as in Figure 8 where the elements of Comp are shown as rounded boxes with their names on top, and certain sites

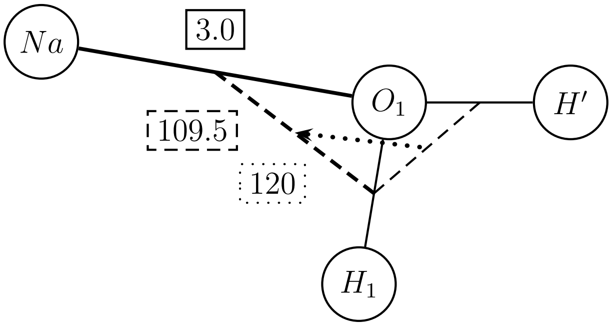


FIGURE 9. Z-system of Na and H_2O

associated to the components are the dots in boxes with the labels. Each component has one indicated site associated to it except for H_2O, which has two different sites. The linkage is the arrow connecting the sites. In this case, all the linkages are directed out from the sites in H_2O.

One method of finding $\mathcal{A}_{r,r'}$ for a particular linkage (r, r') without needing a configuration $R \in \mathcal{B}(\Gamma)$ is to:

(1) create an auxiliary Z-system Λ by tethering the two components;
(2) use the method of Section 1.5 to find a path from r to $\tilde{r}'$, the 2-site extension of the k-site r' for $k = 0, 1, 2$, in the site graph $\mathcal{S}(\Lambda)$;
(3) write the matrix $\mathcal{A}_{r,\tilde{r}'}$, as a product of T-matrices using the tether labels.
(4) if $k = 0, 1$, truncate the matrix to obtain $\mathcal{A}_{r,r'}$.

So we use the method of the undirected site graph discussed in Section 1.7 on the auxiliary Z-system Λ obtained by tethering two components to which r and r' are associated. Using such Z-system, we can construct the coordinate transformation matrix $\mathcal{A}_{r,\tilde{r}'}$ from a 2-site r to a 2-site $\tilde{r}' = (j_0, j_1, j_2)$ which is the 2-site extension of the k-site $r' = (j_0, \ldots, j_k)$ for $k = 0, 1, 2$, resulting from the tethering. To obtain $\mathcal{A}_{r,r'}$ from $\mathcal{A}_{r,\tilde{r}'} = (\mathbf{e}'_0, \mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3)$, we define the column vectors of $\mathcal{A}_{r,r'}$ to be the first 2^k column vectors of $\mathcal{A}_{r,\tilde{r}'}$ so that $\mathcal{A}_{r,r'} = (\mathbf{e}'_0, \ldots, \mathbf{e}'_{2^k-1})$. The labels of the tether simplices determine the matrix $\mathcal{A}_{r,\tilde{r}'}$, and thus $\mathcal{A}_{r,r'}$.

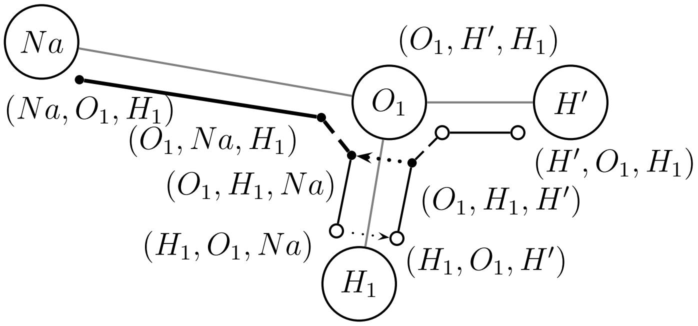


Figure 10. Example of Site Graph

Figure 9 is the tethered Z-system for H_2O and Na. The new bond, angles, and wedges are shown in thicker lines with their labels boxed nearby. We also need a site graph of the tethered Z-system to obtain a sequence of sites in the path from r to r'. An example of a site graph is given in Figure 10, which is the site graph $(\operatorname{vert}\mathcal{S}(\Lambda), \operatorname{edge}\mathcal{S}(\Lambda^*))$ of the Z-system Λ of tethered H_2O and Na. For a set of three atoms, there are six possible sites corresponding to the six different orderings of a set of three elements. If the three atoms are depicted as the three vertices, then the six sites can be depicted as points on the interior of this triangle. There are three pairs of two sites, and each pair can be thought to be related to one of the sides of the triangle, the bonds. However, since only two bonds of a triangle are in the Z-system, by the definition of a site of a Z-system, the two sites that are related to the bond which is not in the Z-system are not sites of the Z-system. So each triangle in a Z-system has four sites, and the total number of the sites in $\operatorname{vert}\mathcal{S}(\Lambda)$ is equal to four times the number of the triangles in the Z-system Λ. Therefore, since Λ has two triangles, there are eight sites shown as small circles in Figure 10. The edges are drawn in different types of lines to represent the different types of T-matrices used to transform between the two sites incident on the edge. The solid line represents an edge of type 1, that is $\{r, r'\} \in \operatorname{edge}_1 \mathcal{S}(\Lambda)$ for two consecutive sites r and r' so that

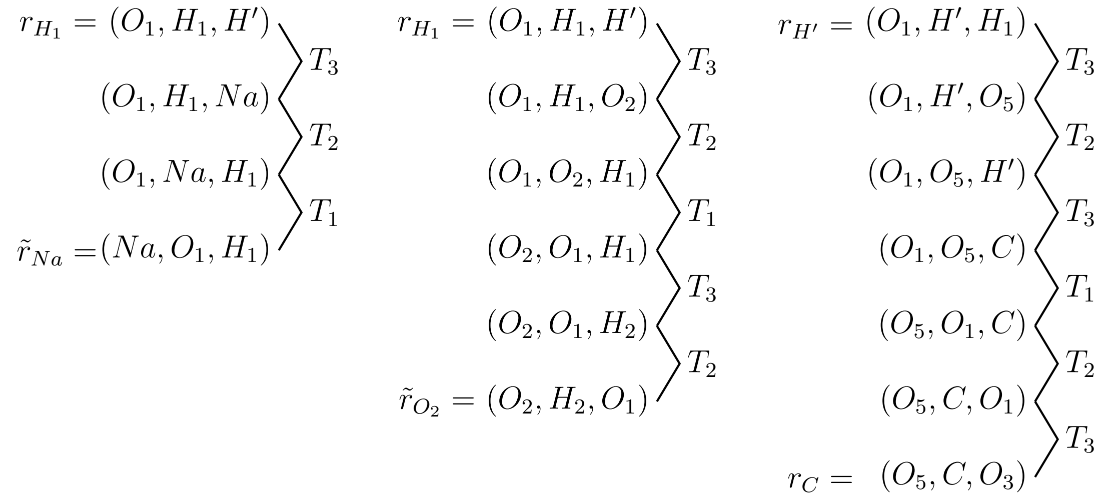


Figure 11. Sequences of Sites and T-Matrices

$E_{r'} = E_r T_1$. If $\{r, r'\} \in \text{edge}_2\, \mathcal{S}(\Lambda)$ so that $E_{r'} = E_r T_2$, then the edge is shown in a dashed line. When $E_{r'} = E_r T_3$, that is, $(r, r') \in \text{edge}_3\, \mathcal{S}(\Lambda^*)$, the edge is drawn in a dotted line with the arrow showing the direction according to the orientation in Λ^3_*.

To find the coordinate transformation matrix $\mathcal{A}_{r_{H_1}, r_{Na}}$ to be assigned to the linkage (r_{H_1}, r_{Na}), the path chosen from (O_1, H_1, H') to (Na, O_1, H_1) is shown by the thick lines for the edges and the darkened circles for the sites in Figure 10. So the sequence of the sites is $((O_1, H_1, H'), (O_1, H_1, Na), (O_1, Na, H_1), (Na, O_1, H_1))$. Comparing the first two consecutive sites in the sequence, (O_1, H_1, H') and (O_1, H_1, Na), the last atom H' is exchanged to Na; so $\{(O_1, H_1, H'), (O_1, H_1, Na)\} \in \text{edge}_3\, \mathcal{S}(\Lambda)$. Also, we have $((O_1, H_1, H'), (O_1, H_1, Na)) \in \text{edge}_3\, \mathcal{S}(\Lambda^*)$ since the sequence order $((O_1, H_1, H'), (O_1, H_1, Na))$ matches the given orientation, $[O_1, H_1, H', Na] = [H', O_1, H_1, Na] \in \Lambda^3_*$. The wedge angle assigned to $[H', O_1, H_1, Na]$ is 120°, so we have $E_{(O_1, H_1, Na)} = E_{(O_1, H_1, H')} T_3(e^{i120^\circ})$ for $e^{i120^\circ} = \cos 120^\circ + i \sin 120^\circ$. If the orientation had been $[Na, O_1, H_1, H']$ with the wedge angle of 120°, we would have used $T_3(e^{i(-120^\circ)}) = T_3(e^{i120^\circ})^{-1}$. For the next two sites from (O_1, H_1, Na) to (O_1, Na, H_1), the last two atoms have been interchanged; thus, $\{(O_1, H_1, Na), (O_1, Na, H_1)\} \in \text{edge}_2\, \mathcal{S}(\Lambda)$. Given the bond angle 109.5° for the triangle $\{O_1, H_1, Na\} \in \Lambda^2$, we have

$E_{(O_1,Na,H_1)} = E_{(O_1,H_1,Na)}T_2(\cos 109.5^\circ)$. The edge incident on the last two consecutive sites is in $\text{edge}_1\ \mathcal{S}(\Lambda)$ since the first two elements had been interchanged from (O_1, Na, H_1) to (Na, O_1, H_1). So $E_{(Na,O_1,H_1)} = E_{(O_1,Na,H_1)}T_1(3.0)$ given the bond length of 3.0 angstrom for the bond $\{Na, O_1\} \in \Lambda^1$. So combining all in order, since $\arccos(-\frac{1}{3}) \approx 109.5^\circ$, we have:

$$E_{(Na,O_1,H_1)} = E_{(O_1,H_1,H')}T_3(e^{i120^\circ})T_2(-\tfrac{1}{3})T_1(3.0).$$

Since $\cos 120^\circ = -\frac{1}{2}$, $\sin 120^\circ = \frac{\sqrt{3}}{2}$, and $\sqrt{1-(-\frac{1}{3})^2} = \frac{2\sqrt{2}}{3}$, we get:

$$\begin{aligned}
&\mathcal{A}_{(O_1,H_1,H'),(Na,O_1,H_1)} \\
&= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ 0 & 0 & \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & \frac{2\sqrt{2}}{3} & 0 \\ 0 & \frac{2\sqrt{2}}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 3.0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & \frac{1}{3} & \frac{2\sqrt{2}}{3} & 0 \\ -\sqrt{2} & -\frac{\sqrt{2}}{3} & -\frac{1}{6} & -\frac{\sqrt{3}}{2} \\ \sqrt{6} & -\frac{\sqrt{6}}{3} & \frac{\sqrt{3}}{6} & -\frac{1}{2} \end{pmatrix}.
\end{aligned}$$

So we now have the coordinate transformation matrix between (O_1, H_1, H') and (Na, O_1, H_1), but we need the matrix $\mathcal{A}_{r_{H_1},r_{Na}}$ between $r_{H_1} = (O_1, H_1, H')$ and $r_{Na} = (Na)$ such that $E_{r_{Na}} = E_{r_{H_1}}\mathcal{A}_{r_{H_1},r_{Na}}$. We know that $E_{r_{Na}} = (\mathbf{e}'_0)$ for $\mathbf{e}'_0 \in \mathbb{R}^3$, is the 0-pose for the 0-site r_{Na}, and that $E_{(Na,O_1,H_1)} = E_{r_{H_1}}\mathcal{A}_{(O_1,H_1,H'),(Na,O_1,H_1)}$ is a 2-pose for the 2-site (Na, O_1, H_1), the extension of the 0-site r_{Na}. By definition of a pose, $\mathbf{e}'_0$ is a position vector of Na, and $\mathbf{e}_0$ in $E_{(Na,O_1,H_1)} = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ is also the position vector of Na, implying that $\mathbf{e}'_0 = \mathbf{e}_0$. To get $\mathbf{e}_0$ from $E_{(Na,O_1,H_1)} = E_{r_{H_1}}\mathcal{A}_{(O_1,H_1,H'),(Na,O_1,H_1)}$, we only need the first column vector of $\mathcal{A}_{(O_1,H_1,H'),(Na,O_1,H_1)}$. Therefore together with

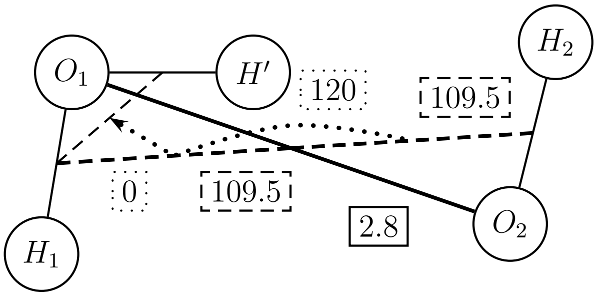


FIGURE 12. Tethered Z-system of OH and H_2O

$\mathbf{e}_0' = \mathbf{e}_0$, we conclude that:

$$E_{r_{Na}} = E_{r_{H_1}} \begin{pmatrix} 1 \\ -1 \\ -\sqrt{2} \\ \sqrt{6} \end{pmatrix}; \text{ thus, } \mathcal{A}_{r_{H_1}, r_{Na}} = \begin{pmatrix} 1 \\ -1 \\ -\sqrt{2} \\ \sqrt{6} \end{pmatrix},$$

and this is the matrix to be assigned to the linkage (r_{H_1}, r_{Na}). Notice that we never actually need either of the poses E_{r_H} or $E_{r_{Na}}$, which would depend on a particular configuration. However, we used the labels of the tethering simplices, which are often of simple geometric interpretation and more accessible to chemical intuition.

Figure 12 is the tethered Z-system for H_2O and OH. The tether labels are chemically natural for a hydrogen bond except for the 120° wedge angle for $[H_1, O_1, O_2, H_2]$, which is arbitrary. We must obtain the sequence of the sites from (O_1, H_1, H') to (O_2, H_2, O_1). One possible sequence is given in the middle column of Figure 11 and on the right of it is the type of T-matrix used to transform the coordinates of each site to those of the next. This sequence can be derived naturally by considering the starting site (O_1, H_1, H'), the ending site (O_2, H_2, O_1), and the auxiliary Z-system shown in Figure 12; it is not necessary to draw the site graph. By substituting the given value appropriately into each T-matrix, since $e^{i0^\circ} = \cos 0^\circ + i \sin 0^\circ$ and

$e^{i(-120^\circ)} = \cos(-120^\circ) + i\sin(-120^\circ)$, we get:

$$\begin{aligned}
E_{(O_2,H_2,O_1)} &= E_{(O_1,H_1,H')} T_3(e^{i0^\circ}) T_2(-\tfrac{1}{3}) T_1(2.8) T_3(e^{i(-120^\circ)}) T_2(-\tfrac{1}{3}) \\
&= E_{(O_1,H_1,H')} \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{14}{15} & -\frac{5}{9} & \frac{\sqrt{2}}{9} & \frac{-\sqrt{6}}{3} \\ \frac{28\sqrt{2}}{15} & \frac{\sqrt{2}}{9} & \frac{17}{18} & -\frac{\sqrt{3}}{6} \\ 0 & \frac{-\sqrt{6}}{3} & -\frac{\sqrt{3}}{6} & \frac{1}{2} \end{pmatrix}.
\end{aligned}$$

Notice that for the second T_3 matrix, we use $e^{i(-120^\circ)}$ instead of e^{i120°. This is because the sequence order $((O_2, O_1, H_1), (O_2, O_1, H_2))$ matches to the orientation $[H_1, O_2, O_1, H_2]$ which is opposite from the given orientation $[H_1, O_1, O_2, H_2] = 120^\circ$. By the definition of a pose with the order of the atoms in (O_2, H_2, O_1) compared to $r_{O_2} = (O_2, H_2)$, the first two column vectors of $E_{(O_2,H_2,O_1)}$ are equal to the two column vectors of $E_{r_{O_2}}$, that is, for the 2-pose $E_{(O_2,H_2,O_1)} = (\mathbf{e}_0, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and the 1-pose $E_{r_{O_2}} = (\mathbf{e}'_0, \mathbf{e}'_1)$, $\mathbf{e}'_0 = \mathbf{e}_0$ and $\mathbf{e}'_1 = \mathbf{e}_1$. Then,

$$E_{r_{O_2}} = E_{(O_2,H_2,O_1)} \begin{pmatrix} 1 & 0 \\ -\frac{14}{15} & -\frac{5}{9} \\ \frac{28\sqrt{2}}{15} & \frac{\sqrt{2}}{9} \\ 0 & -\frac{\sqrt{6}}{3} \end{pmatrix}, \text{ so } \mathcal{A}_{r_{H_1}, r_{O_2}} = \begin{pmatrix} 1 & 0 \\ -\frac{14}{15} & -\frac{5}{9} \\ \frac{28\sqrt{2}}{15} & \frac{\sqrt{2}}{9} \\ 0 & -\frac{\sqrt{6}}{3} \end{pmatrix},$$

and $\mathcal{A}_{r_{H_1}, r_{O_2}}$ is assigned to the linkage (r_{H_1}, r_{O_2}).

For the linkage $(r_{H'}, r_C)$, we consider Figure 13 for the tethered Z-system of H_2O and Carbonic Acid. Naturally, the two atoms that form a hydrogen bond and the atoms covalently bonded to those are collinear or close to being on a straight line. So we choose $\{O_1, H', O_5\}$ to be the new angle instead of $\{O_1, H_1, O_5\}$, which would have been close to 0° or 180° so that the site is $r_{H'}$ instead of r_{H_1}. Using the sequence of sites as shown in the right column of Figure 11 with the labels given in the Z-system in Figure 13, we find the coordinate transformation matrix for $(r_{H'}, r_C)$. Since

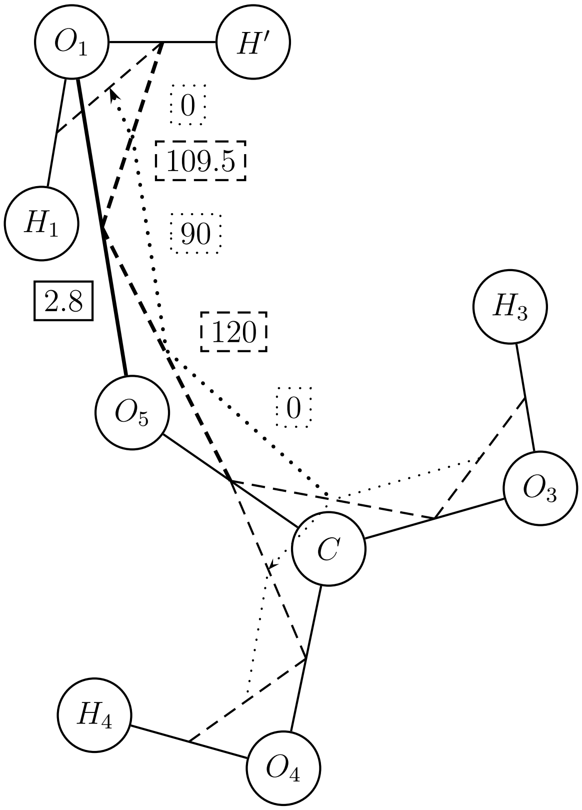


FIGURE 13. Z-system of Carbonic Acid and H_2O

$e^{i90^\circ} = \cos 90^\circ + i \sin 90^\circ$,

$$\begin{aligned} E_{r_C} &= E_{r_{H'}} T_3(e^{i0^\circ}) T_2(-\tfrac{1}{3}) T_3(e^{i90^\circ}) T_1(2.8) T_2(-\tfrac{1}{2}) T_3(e^{i0^\circ}) \\ &= E_{r_{H'}} \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{14}{15} & -\frac{1}{6} & \frac{\sqrt{3}}{6} & -\frac{2\sqrt{2}}{3} \\ \frac{28\sqrt{2}}{15} & \frac{\sqrt{2}}{3} & -\frac{\sqrt{6}}{3} & -\frac{1}{3} \\ 0 & -\frac{\sqrt{3}}{2} & -\frac{1}{2} & 0 \end{pmatrix}. \end{aligned}$$

Since r_C is a 2-pose, we need no further comment but to conclude that

$$\mathcal{A}_{r_{H'}, r_C} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{14}{15} & -\frac{1}{6} & \frac{\sqrt{3}}{6} & -\frac{2\sqrt{2}}{3} \\ \frac{28\sqrt{2}}{15} & \frac{\sqrt{2}}{3} & -\frac{\sqrt{6}}{3} & -\frac{1}{3} \\ 0 & -\frac{\sqrt{3}}{2} & -\frac{1}{2} & 0 \end{pmatrix}$$

is the matrix assigned to the linkage $(r_{H'}, r_C)$.

With this, we finish the labeling of the linkages and hence complete the labeling of the GZ-system for the molecular system of sodium bicarbonate dissolved in water. The auxiliary tethered Z-systems, such as Λ^*, are discarded once we find the coordinate transformation matrix for each linkage.

The concept of the GZ-system and its way of defining the conformation becomes necessary and useful in building a model for a large system of biological molecules. The necessity became clearer as we tried to build a model of a system with many parts. In the study of a biological molecular system, the system often involves many molecules that are relatively free in motions. If we were to tether all those molecules into one large Z-system, some angles might be close or equal to 0° or 180°, the singularity points of Z-system coordinates. (This problem with angles of 0° or 180° does not occur within components because strong chemical forces keep the values of those angles well removed from the singular values.) As mentioned before, this can lead to numerical problems in simulations, or some undefined wedge angles in certain special confomrations. However the use of $\mathcal{A}$-matrices and multi-component GZ-systems frees us from this problem since no entry of any of the $\mathcal{A}$-matrices depends on the conformation in a nonsmooth manner.

Chapter 2

Light-Harvesting Complex

2.1. Biological Setting

Plants are known to provide oxygen in the atmosphere through the process called **photosynthesis**. In this process, plants use the energy of sunlight to produce oxygen and carbohydrates, used as an energy source in plants and animals, from carbon dioxide, CO_2, and water. There are some bacteria that also carry out photosynthesis. The **photosynthetic unit (PSU)**, of which the **light-harvesting complex (LHC)** is a part, is where the first step of photosynthesis takes place [22].

2.1.1. The Photosynthetic Unit. The PSU lies in the lipid bilayer membrane of the bacterial cell. It consists of the **reaction center (RC)** and the LHC made of proteins and of light absorbing pigments called **bacteriochlorophylls (BCLs)** and **carotenoids** [12, 14, 27].

In purple bacteria, the PSU usually contains two types of LHCs, commonly referred to as LH-I and LH-II. BCLs in LH-I absorb light at a wavelength of about 875 nm, and so are sometimes referred to as B875. LH-II contains two types of BCLs, B850 and B800, that absorb light at wavelengths of around 850 nm and 800 nm, respectively [12]. Some species have another type of LHC which is referred to as LH-III whose BCLs absorb light at wavelengths 820 nm and 800 nm [12, 27]. The carotenoids assist the BCLs by absorbing light at a wavelength around 500 nm and transfering the energy to the BCLs.

The RC is directly surrounded by a circular LH-I. The number of LH-IIs depends on the growth conditions and the light intensity, but about eight to ten ring-shaped LH-IIs surround LH-I [12,14]. There are two theories regarding to how LH-I, RC, and bc$_1$, which is another system that takes a part in photosynthesis, are arranged [13], but that ambiguity will not affect our study of LH-II.

2.1.2. LH-II.

LH-II is a **transmembrane** complex; that is, LH-II is longer than the thickness of the cell membrane so that the complex crosses the membrane with one part sticking out on one side of the membrane (called the **cytoplasm** side) and another part on the other side (called the **periplasm** side).

LH-II is made from multiple identical subunits called **protomer complexes (PCs)**. The number of PCs in an LH-II depends on the species [14]; usually eight or nine PCs are assembled to form the ring of LH-II (see Appendix A, Figure 35). A PC consists of one of each α- and β-**apoprotein** and three BCLs and usually one or two carotenoids [23] (Figure 36). A pair of α- and β-apoproteins is called a **heterodimer**, and both apoproteins have **α-helical** structure. The two helices of the heterodimer are parallel to each other and almost perpendicular to the membrane plane with the α-apoprotein on the inside and β-apoprotein on the outside of the LH-II ring. To each heterodimer, three BCLs are noncovalently bonded: two B850 BCLs and one B800 BCL. The head-like flat surface of the two B850 BCLs are situated perpendicular to the membrane plane closer to the periplasm side. They are placed in between the α- and the β-apoproteins in such a way that in the ring of LH-II, one B850 BCL in one PC overlaps with B850 BCL of the neighboring PC on one side and the other B850 BCL overlaps with B850 BCL of the PC on the other side. So we have a ring of B850 BCLs within the ring of LH-II (Figure 38). The flat surface of the B800 BCL lies under the tail of one of the B850 BCLs between two neighboring PCs, almost parallel to the membrane plane close to the cytoplasm side [23] (Figure 39).

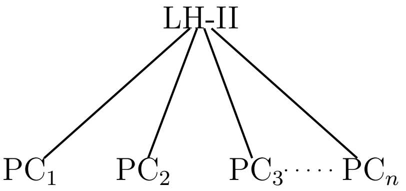


Figure 14. Hierarchy Tree of LH-II

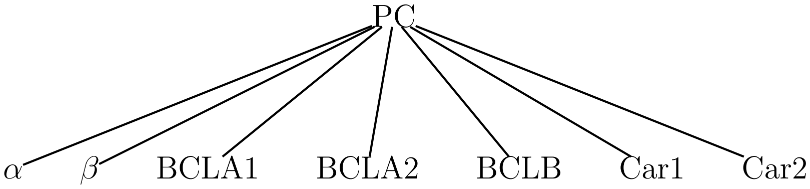


Figure 15. Hierarchy Tree of PC

The one carotenoid which is surely present in a PC runs from the cytoplasm side of the complex, sliding in between the two transmembrane α-helices of the heterodimer, to the periplasm side (Figure 40). The other carotenoid, which is only present in on the average of one of every two PCs, occupies a pocket between two adjacent PCs (Figure 41). The exact positioning of this carotenoid is uncertain, and it may be disordered ***in vivo*** [23].

The way LH-II is organized in its structure suggests a certain hierarchy tree: LH-II is the top level and an individual PCs on the next level (Figure 14). Each PC branches out to the another level having α-apoprotein, β-apoprotein, BCLs, and carotenoids (Figure 15). We will see later how α- and β-apoproteins are extended in Section 2.2.2.

2.2. Chemical Nature of Building Blocks

2.2.1. Membrane Lipids. The cell membrane of purple bacteria, where the PSU is located, is surrounded by water. By their chemical nature, water molecules are always seeking to interact with other molecules to form **hydrogen bonds**. When a water molecule is next to some molecules that do not form such bonds, the water molecule cannot move as freely as it would around other water molecules; thus, the free energy of that state is higher.

The cell membrane is made of two layers of lipids. Lipids consist of two parts: a hydrophilic head and hydrophobic tail. The term **hydrophilic** means that this part of a molecule tends to form a hydrogen bonds with water molecules. On the other hand, a **hydrophobic** part of a molecule does not interact with water molecules. Since molecules tend to stay in the lowest free energy configuration, the hydrophilic head faces and interacts with the surrounding water, and the hydrophobic tail stays away from the water molecules as much as possible. As a result, the hydrophilic heads of the lipids of each layer face the water on each side, and the hydrophobic tails face each other, staying away from the water. So both surfaces of the membrane are the hydrophilic part of the lipids and the hydrophobic part is in between them.

Lipids are a diverse group of molecules. Lipids that form the cell membrane in organisms often have two hydrophobic tails resulting in a cylindrical structure which can easily pack in parallel to form a sheet of bilayer membrane. Glycerophospholipids (or *phosphoglycerides*) have such a feature and are also a major class of naturally occurring phospholipids, lipids which have phosphate in the head groups [22]. These lipids make up a significant part of the membrane lipids throughout living organisms. As the name suggests, glycerophospholipids are derivatives of glycerol, and each has one **polar** side chain bonded to phosphate which is bonded to glycerol for the head group and two hydrocarbon tails. The types of glycerophospholipids vary depending on the side chains of the hydrophilic head groups. According to the table in [22], 83

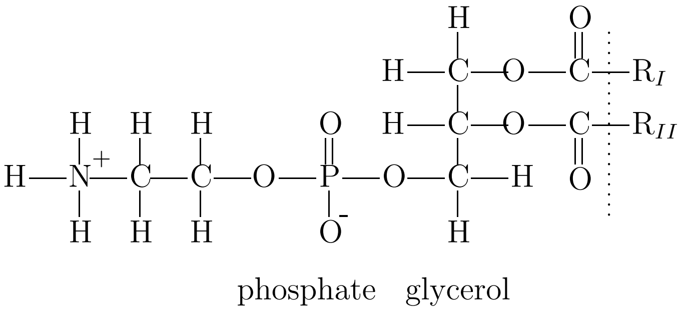
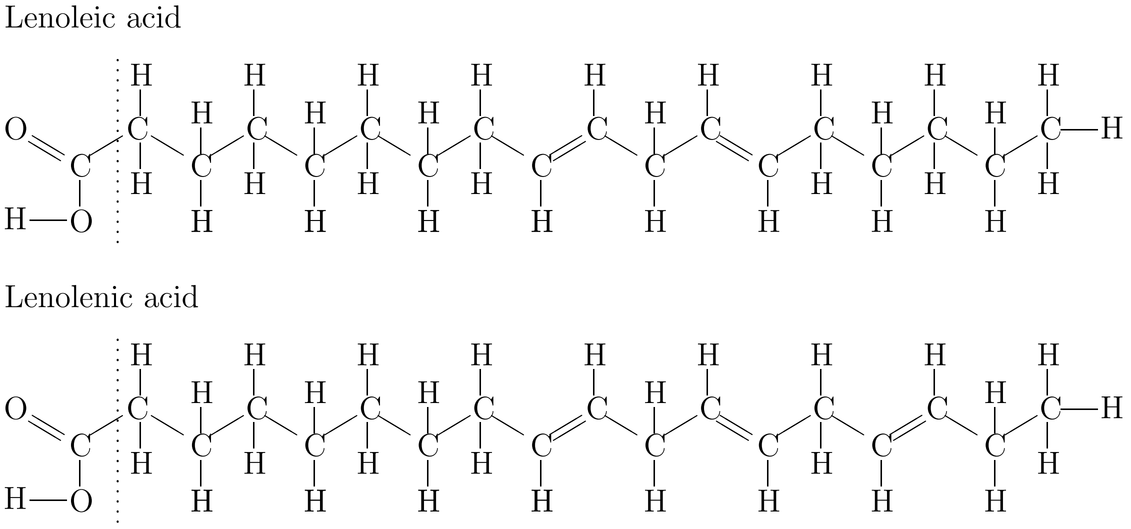


Figure 16. Phosphatidylethanolamine


Lenoleic acid

Lenolenic acid

Figure 17. Examples of Fatty Acids

percent of the cell membrane of a common bacterium *Escherichia (E.) coli* consists of glycerophospholipids: 65 percent of phosphatidylethanolamine and 18 percent of phosphatidylglycerol. Figure 16 shows the chemical structure of the polar head part of phosphatidylethanolamine with R_I and R_{II} representing two hydrocarbon tails. These two tails are derived from naturally occurring fatty acids, one of the simplest type of lipids by themselves. Figure 17 shows two examples of fatty acids, lenoleic and lenolenic acids.

2.2.2. Protein. The basic building blocks of any protein are the **amino acids**. Each amino acid can be thought to have two parts: the **backbone** and the side chain. An individual amino acid has three hydrogen atoms (H) attatched to the nitrogen

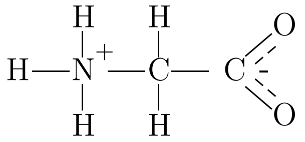


FIGURE 18. Glycine Amino Acid (Backbone with an H side chain)

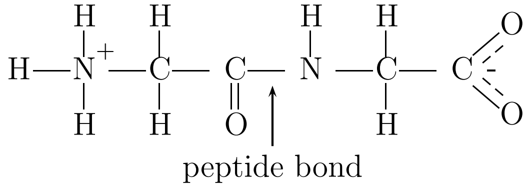


FIGURE 19. A Chain of Two Glycine Amino Acids

atom (N) of the backbone and two oxygen atoms (O) to the carbon atom (C) of the backbone (Figure 18). Since a nitrogen atom bonded to three hydrogens has a positive charge and a carbon atom bonded to two oxygens has a negative charge, they are attracted to each other and can react to form a bond between the two amino acids, producing a water molecule as a byproduct (Figure 19). When multiple amino acids react in this way, a protein chain is formed. A short chain of amino acids is called a **peptide**, and the bond between the two amino acids is called the **peptide bond**.

Sometimes "backbone" may refer only to the chain of nitrogen and carbon atoms. The tip of the backbone chain which ends with an **amino group**, NH_3^+, is called the **N terminus**, and the other end which ends with a **carboxyl group**, COO^-, is called the **C terminus**. In some proteins, these terminii are modified so that, for example, instead of three hydrogen atoms attatched to the nitrogen on the N teminus, one might have a carboxyl group attached to the nitrogen. Amino acids are numbered in the protein counting from the N terminus, and a part of an amino acid which is left after forming a peptide is called **residue**.

Alanine	ALA	A	Leucine	LEU	L
Arginine	ARG	R	Lysine	LYS	K
Asparagine	ASN	N	Methionine	MET	M
Aspartic acid	ASP	D	Phenylalanine	PHE	F
Cysteine	CYS	C	Proline	PRO	P
Glutamic acid	GLU	E	Serine	SER	S
Glutamine	GLN	Q	Threonine	THR	T
Glycine	GLY	G	Tryptophan	TRP	W
Histidine	HIS	H	Tyrosine	TYR	Y
Isoleucine	ILE	I	Valine	VAL	V

Table 2. Amino Acid Names and Abbreviations

Not only in LH-II but in any biomolecule, hydrogen (H) bonds play a major role in stabilizing the structure of the molecule. Especially in proteins with α-helical structure such as the α- and β-apoproteins of LHCs, the hydrogen atom (H) on the nitrogen atom (N) of the backbone forms a hydrogen bond with the oxygen atom (O) double bonded to the carbon (C) of the backbone. All together these form a string of hydrogen bonds vertically along the helix, and they hold the molecule in its helical structure. In a system of molecules like LH-II, hydrogen bonds also help to hold the different molecules together.

There are 20 different biological amino acids. The chemical structure of a glycine amino acid is shown in Figure 18. For the other amino acids, we replace one of the hydrogen atoms from the middle carbon atom, called C^α, with the appropriate side chain shown in Figures 20, 21, and 22. (In those figures, a vertex with no atom symbol is a carbon atom by convention.) The backbone part of amino acids is the same for all amino acids with some modification for proline as shown in Figure 20 with its modified backbone.

Table 2 lists the 20 amino acids with their three-letter abbreviations in the second column and one-letter abbreviations in the third column. These amino acids, all except glycine, are grouped into three different categories according to their chemical nature. The hydrophobic amino acids, shown in Figure 20, are the amino acids that

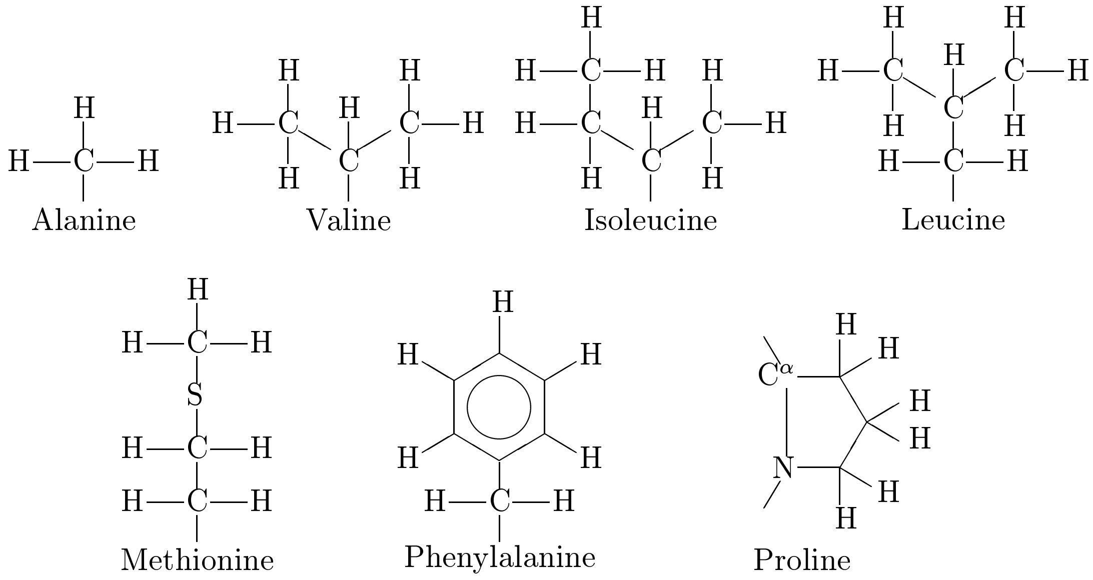
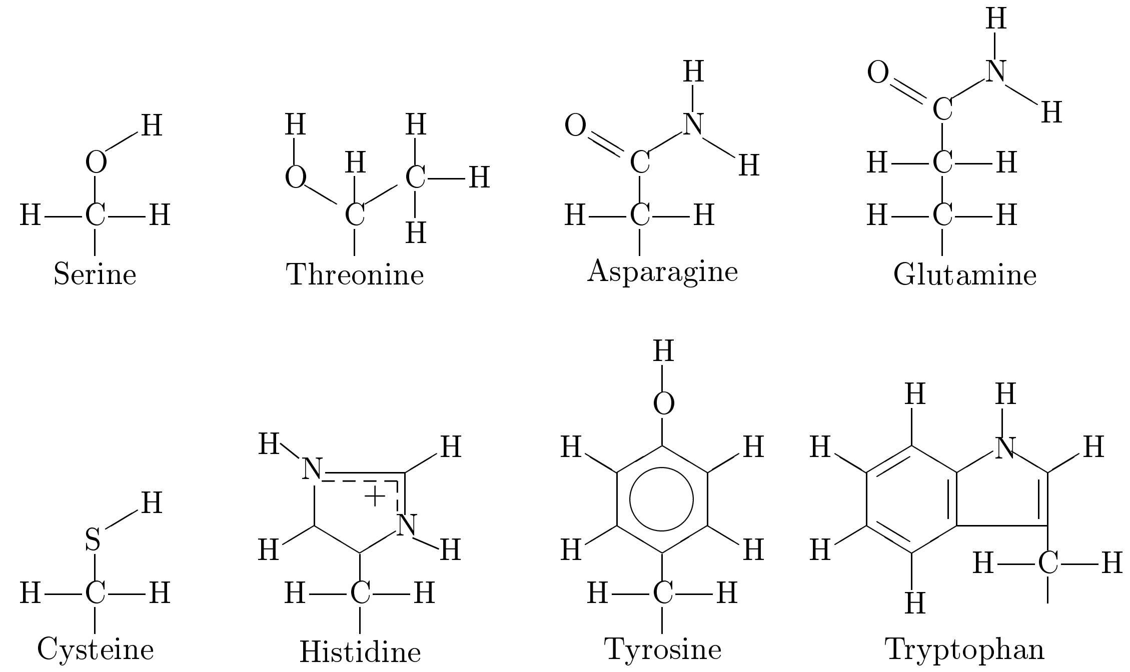


Figure 20. Hydrophobic Amino Acids

are hydrophobic as the name suggests, so they do not easily react with other molecules for the most part. The bonding between carbon and hydrogen atoms has this nature since the charge is shared equally between the two types of atoms. Sulfurs (S) can form hydrogen bonds but not strong ones.

In other bonds such as a bond between an oxygen atom (O) and a hydrogen atom (H) or a nitrogen atom (N) and a hydrogen atom (H), the two atoms do not share the charge equally. As a result, one atom (O or N) is slightly negatively charged and the other (H) is slightly positively charged. Such a bond is said to be polar, and polar amino acids have such bonds (Figure 21). Among those polar amino acids, histidine is a unique amino acid. In a neutral state, either nitrogen in the five-membered ring (a ring of five covalent bonds formed between five atoms), but not both, will be bonded to hydrogen. It is **protonated** if both nitrogen atoms are bonded to hydrogen atoms and is positively charged as a whole.

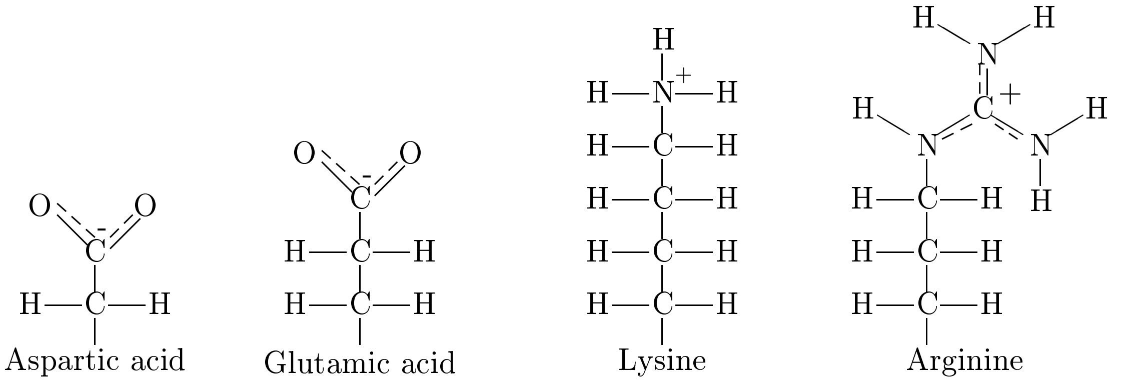
Another group of amino acids, the charged amino acids take different forms depending on the availability of hydrogen atoms in the surrounding environment. The forms shown in Figure 22 are the most likely ones for these amino acids at pH=7. If



Serine Threonine Asparagine Glutamine

Cysteine Histidine Tyrosine Tryptophan

Figure 21. Polar Amino Acids



Aspartic acid Glutamic acid Lysine Arginine

Figure 22. Charged Amino Acids

the environment is more **acidic** (such as pH=4), which means that more hydrogen ions, H^+, are present, then the negatively charged oxygen atom in aspartic acid or glutamic acid are likely to be bonded to another hydrogen atom. If the environment is more **basic** (such as pH=10) so that fewer hydrogen ions are available than when pH=7, then the positively charged nitrogen atom in lysine or arginine will release one hydrogen ion and become neutral.

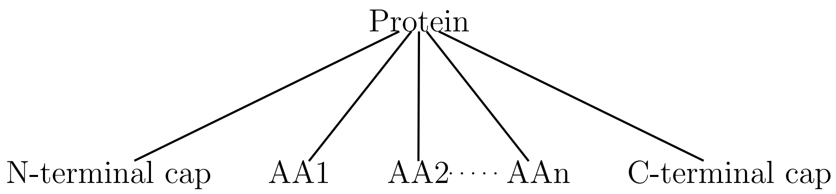


Figure 23. Hierarchy Tree for Protein

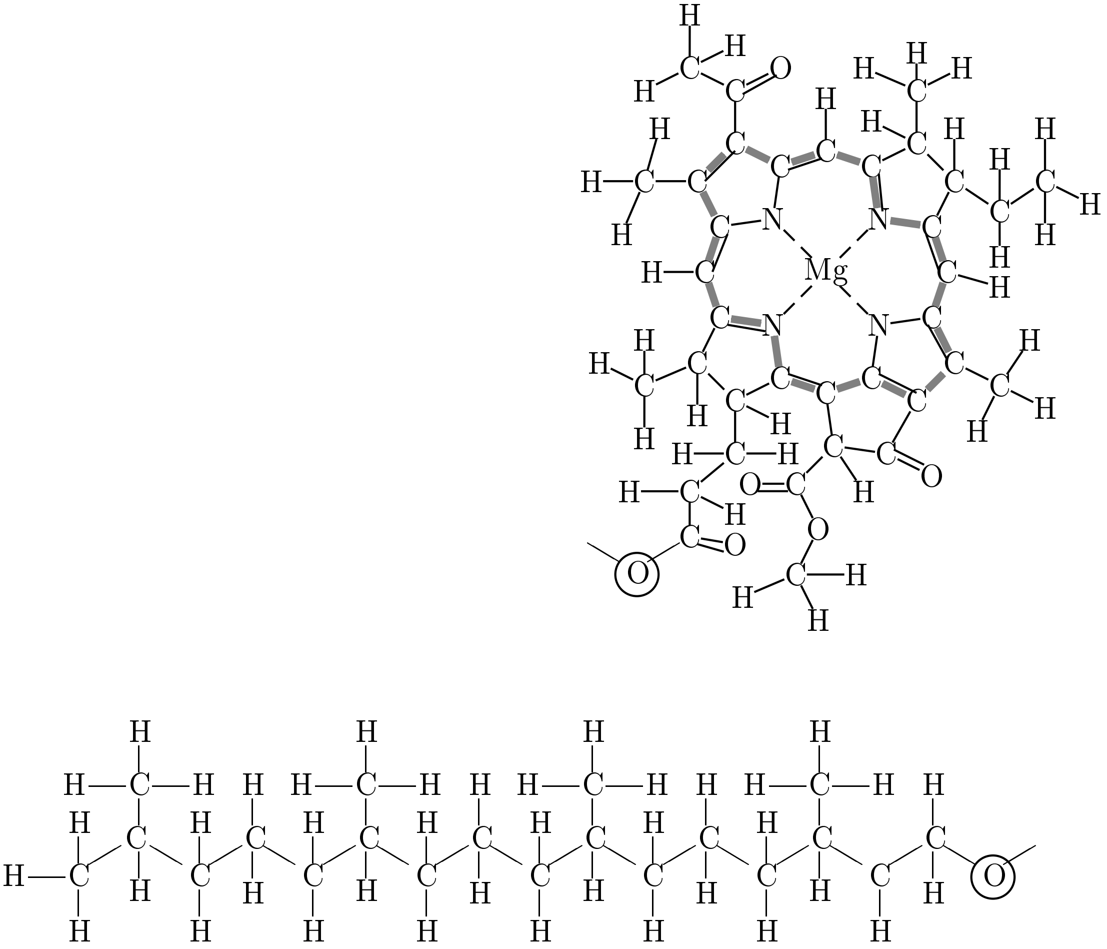


Figure 24. Chemical Structure of BCL a

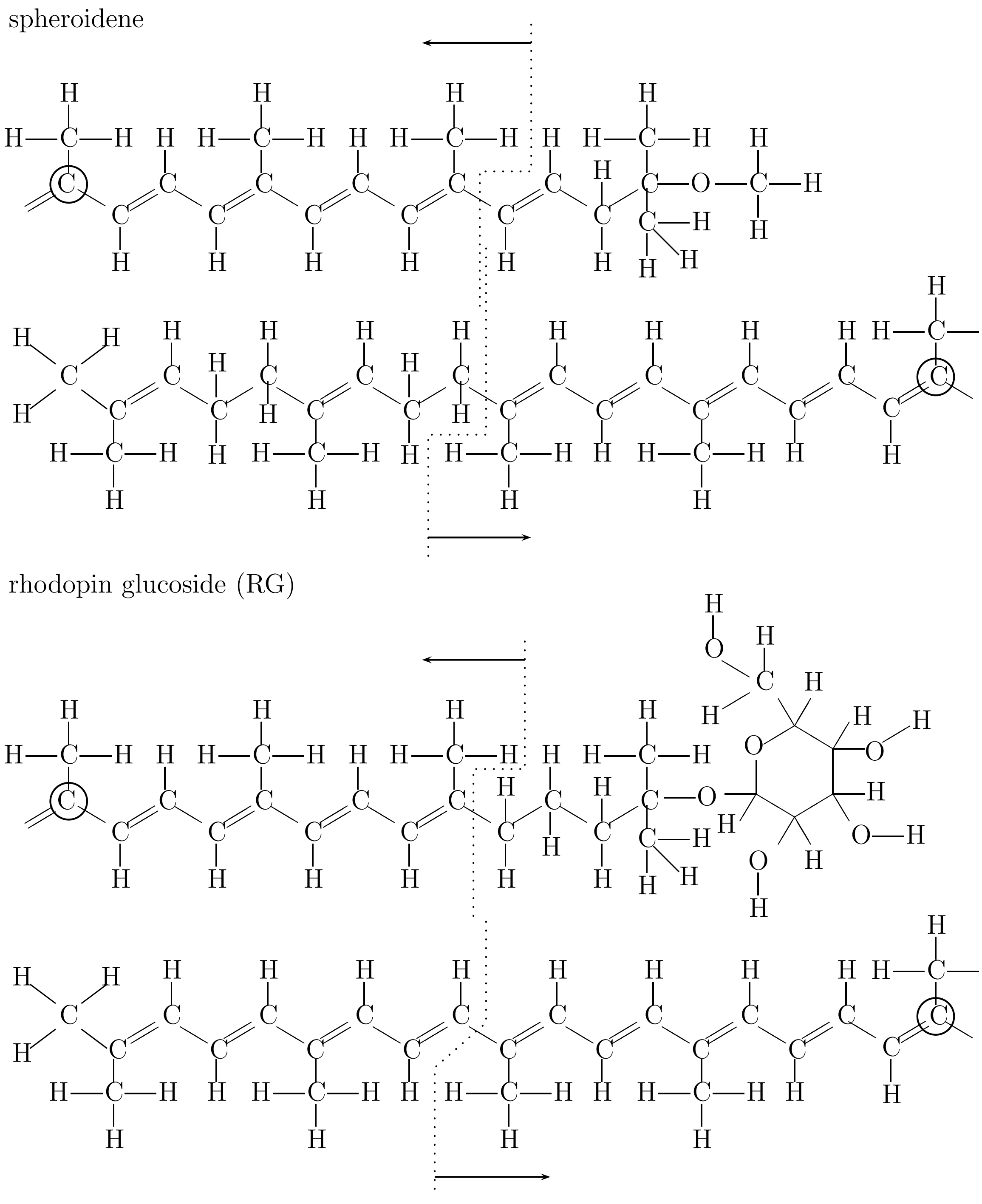
Since α- and β-apoproteins are protein chains made of amino acids, we can extend the earlier hierarchy tree in Section 2.1.2 from them to have another level and have amino acids under both apoproteins. Under each amino acid, we have another level for the atoms (Figure 23).

2.2.3. Bacteriochlorophyll. A bacteriochlorophyll has a head-like flat surface part and a tail-like hydrocarbon chain. The chemical structure of BCL a is shown in Figure 24. Some species have other BCLs that are chemically different such as BCL b in *Rps. virdis* [14], but since the species on which we will focus only contain BCL a, here BCL will refer to BCL a. The shaded region of the head part in Figure 24 shows the **conjugated bonds**, a chain of alternating single bonds and double bonds. This alternation of single and double bonds allows the pigment molecules such as BCL to interact with the light energy and to absorb it and to pass it on to RC.

In the middle of the head-like structure of BCL is a magnesium atom. Magnesium is a metal and therefore is easily bonded to other atoms. One B850 BCL in a PC of LH-II is **ligated** to the α-apoprotein, sometimes referred to as B850a, and another one is ligated to β-apoprotein, referred to as B850b. The tail part of BCL is a single-bonded hydrocarbon chain; thus, that part is flexible. Since the hydrocarbon part is hydrophobic, the tail part tends to be away from water.

2.2.4. Carotenoid. Carotenoids also have a hydrophobic tail-like structure, but unlike BCL's tail, it contains conjugated bonds. They absorb the light energy through this conjugated tail. The alternation of single and double bonds more or less fixes this part to be flat, allowing it to fit between the two transmembrane helices of the heterodimer. Some carotenoids do have only single bonds in some parts of their hydrocarbon tails so that the shape can be flexible in those parts.

The types of carotenoids contained in LH-II differ from species to species, and some species may have several types of carotenoids. LH-II from *Rs. molischianum* is known to have lycopene as its carotenoid [12]. Spheroidene and spheroidenone are two chemical structure-wise similar carotenoids found in *Rb. sphaeroides*, although spheroidene, shown at the top of Figure 25 is more common in LH-II [15]. *Rps. acidophila* has rhodopin glucoside (RG), shown at the bottom of Figure 25, rhodopin,



spheroidene

rhodopin glucoside (RG)

Figure 25. Chemical Structures of Carotenoids

and lycopene as possible types of carotenoids in its LH-II [23]. The chemical structures of some other carotenoids can be found in [30].

Chapter 3

Homology Modeling

3.1. Homology Modeling Overview

As we have discussed earlier, proteins are made of amino acids. When the sequence of amino acids in one protein chain is **homologous**, or alike, to that of another protein chain, especially when the functions of those proteins are similar, we often assume that their structures resemble one another. In **homology modeling**, based on this assumption, we build a model of one protein chain (target) for which the structure is not known from the known structure of another protein chain (template). First, templates are chosen for a target by comparing sequences of amino acids. The choice of templates and how their sequences should be aligned with the sequence of the target protein chain are fine tuned by considering other factors such as biological function and **secondary structure**, the structure of the backbone. Then the sequence of amino acids of the target is put onto the structure of the templates.

(For a more detailed description of homology modeling, the reader is referred to [17], and also to [14] for a particular example of how homology modeling is done.)

- **Goal**: To build a model for LH-II in *Rb. sphaeroides.*
- **Principles**:
  - Attempt to keep the core structure invariant.
  - Maintain perfect symmetry about the central axis of LH-II.
  - Build the model with Z-system internal coordinates.
  - Keep steric consistency
  - Attempt to have all residues in their rotamers.
  - Attempt to achieve all of the above by simple adjustments.

FIGURE 26. The Principles Governing the Experiment

3.2. Goal and Principles

The goal of our experiment is to build a structure of LH-II from *Rb. sphaeroides* using the available structures of LH-II from other species, mainly *Rps. acidophila*, as the templates. We aim to achieve this by following some principles. Our objective is to see if we can build a model of a protein complex using the known structure of another protein complex keeping intact the structure for the major parts. So the first principle is to attempt to keep the protein backbone structures, the pigment conformations, and the relative positions of the components within a PC and the relative placements of the PCs within the LH-II complex invariant. In regard to the placements of PCs within LH-II, we require the PCs to maintain perfect symmetry about the central axis of the LH-II ring, that is, all PCs should have identical conformations. In building a model, our fundamental principle is to build the structure entirely in Z-system internal coordinates in order to facilitate adjustability.

Also for the resulting structure to be reasonable, the choice of the conformations of all residues must be such that no two non-bonded atoms are too close to each other in space. We attempt to accomplish this using a rotamer conformation (discussed in Section 4.1.2) for each side chain, with the rotameric conformations closest to the conformation observed in the templates for all **conserved** residues, that is the residues whose side chains are identical between the species. We also would like to achieve these principles only by adjusting a small number of internal coordinates one by one by hand.

3.3. LH-II FROM *Rb. sphaeroides*

3.3.1. Target and Templates. The detailed structure of the PSU of purple bacteria is not exactly known, but the architecture of all LHCs is remarkably similar, and also the function of the LHC is known. [12] reports that LH-II from *Rb. sphaeroides* was modeled successfully as having nine PCs using LH-II from *Rps. acidophila* as the template. So we chose LH-II from *Rb. sphaeroides* as our target and LH-II from *Rps. acidophila* as our template. Since we focus on the structure of LH-II, we assume the α- and β-apoproteins discussed here refer to those of LH-II, unless otherwise noted. Also we abbreviate *Rb. sphaeroides* as *sph.*

The amino acid sequences of α- and β-apoproteins from *sph.* were obtained from Swiss-Prot [2], a curated protein sequence database which strives to provide a high level of annotations. The sequences for the apoproteins from *Rps. acidophila* were obtained from the file named 1NKZ in the Protein Data Bank (PDB) [1], the repository for the processing and distribution of 3-D macromolecular structure data. We aligned the sequences simply by hand, one from *sph.* and another from *Rps. acidophila*, for each apoprotein. For different possible alignments (assuming no gaps), we counted the number of residues for which the corresponding amino acids were identical. We used the particular alignment that gave the highest numbers of identical residues, which turned out to agree with the alignments given in [14] for both apoproteins. The sequences of *Rps. acidophila* given in the PDB do not match exactly with the sequences in the multi-**sequence alignment** (an alignment of multiple amino acid sequences) in [14] since they are from different strains of *Rps. acidophila.* We use *acid.* as an abbreviation for the strain of *Rps. acidophila* whose sequences are given in the PDB and *aci** for the strain in the multi-sequence alignment in [14].

Having done the alignment by hand as shown in Figure 27, we observe that the sequence of *acid.* for α-apoprotein has one less residue than that of *sph.* on the C terminus. (Thus, the last residue of *sph.* α-apoprotein is not included in our structure.)

```
                                     α-apoprotein
                10          20          30          40          50
sph.    XTNGKIWLVV  KPTVGVPLFL  SAAVIASVII  HAAVLTTTTW  LPAYYQGSAA  VAAE
acid.   XNQGKIWTVV  NPAIGIPALL  GSVTVIAILV  HLAILSHTTW  FPAYWQGGVK  KAA-

                                     β-apoprotein
                 6          16          26          36          46
moli.   ----AERSLS  GLTEEEAIAV  HDQFKTTFSA  FIILAAVAHV  LVWVWKPWF-
                10          20          30          40          50
sph.    TDDLNKVWPS  GLTVAEAEEV  HKQLILGTRV  FGGMALIAHF  LAAAATPWLG
                 1          11          21          31          41
acid.   ---------A  TLTAEQSEEL  HKYVIDGTRV  FLGLALVAHF  LAFSATPWLH
```

Figure 27. The Sequence Alignment

For the β-apoprotein, the sequence from the *acid.* is nine residues shorter than that from *sph.* on the N terminus. So an attempt was made to add these residues of the β-apoprotein on the N terminus side from another species to predict the backbone structure of the β-apoprotein from *sph.* Because of the high homology—especially the residues on the N terminus side—and the availability of the atomic-level structure (although missing the structure for the two residues on the N terminus), the β-apoprotein from *Rs. molischianum*, whose abbreviation will be *moli.*, was chosen. The file name in the PDB is 1LGH.pdb.

3.3.2. Transition of the Templates. The sequence of *moli.* β-apoprotein conserves the glycine residue 7, G7.B.*moli.*, which is highly conserved across the species including *sph.* [14]. (For the residues, we will use the notation AAnum.apo.species where AA is the one letter abbreviation of an amino acid name, num is the residue number, apo is either A for α-apoprotein or B for β-apoprotein, and the species abbreviation.) Highly conserved means that the amino acids of the corresponding positions in the sequences of amino acids in LHCs from other species are mostly also glycines, but this glycine residue is not conserved in the β-apoprotein from *acid.* G7.B.*moli.* corresponds to G11.B.*sph.* and to T2.B.*acid.* The glutamic acid of residue 16, E16.B.*sph.*, and alanine of residue 17, A17.B.*sph.*, are conserved between β-apoproteins from *sph.* and from *moli.* These two residues are also conserved in ten out of the twelve sequences of the β-apoproteins aligned in the multi-sequence alignment in [14]. Among those ten sequences is the sequence of β-apoprotein from *aci**. In *acid.* β-apoprotein, those two residues, E16.B.*sph.* and A17.B.*sph.*, are not conserved; their corresponding residues are Q7.B.*acid.* and S8.B.acid., respectively. However, the following two residues, E18.B.*sph.* and E19.B.*sph.*, are conserved with *acid.*. Furthermore, the latter residue, E19.B.*sph.*, is highly conserved according to the multi-sequence alignment in [14]. Interestingly, the only non-conserving sequence is the sequence from *moli.* where the corresponding residue is A15.B.*moli.* Considering these facts, we decided to use *moli.* as the template for residues 7 to 17 of *sph.*, and *acid.* as the template for residues 18 to 50. Since we do not have any structural information on the residues 1 to 6 of the β-apoprotein for *sph.*, we do not include these residues in our structure.

3.3.3. Sequence Homology. With the alignment shown in Figure 27, the sequences of α-apoproteins from *sph.* and *acid.* have 24 identical residues. Among the residues that are not identical, there are 19 residues that the corresponding amino acids are in the same group. So 43 residues out of the 53 possible are homologous.

For the sequences of β-apoproteins from *sph.* and *acid.*, 25 residues are identical and 32 are homologous out of the 41 possible. Between the β-apoproteins from *sph.* and *moli.*, 18 are identical and 29 are homologous out of the 45 possible. By using *moli.* for the residues 7 to 17 of *sph.* and *acid.* for the residues 18 to 50, we count 29 identical residues and 37 homologous out of the 46 possible.

Chapter 4

Building a Model

4.1. The Z-system of A Protein Chain

As discussed in Section 2.2.2, the building blocks of proteins, amino acids, can be thought of as having two parts: the backbone and the side chain. We utilize the characteristic of backbone repetition by making the Z-system for the backbone separately from those of side chains.

Each atom of an amino acid has a name by which it is known according to the conventions in [21]. In the backbone of an amino acid, the nitrogen atom is called N, the carbon atom bonded to N is C^{α}, and the carbon atom bonded to C^{α} is C' or just C. Sometimes C^{α} may be referred to as CA with A standing for α. The non-hydrogen atoms in the side chains are given names with letters of the Greek alphabet following α. So a carbon atom bonded to C^{α} is named C^{β}, or CB, and the following would be C^{γ}, or CG, and C^{δ}, or CD, and so on. Naturally, if oxygen is bonded to C^{β}, for example, it would be named O^{γ}, or OG.

As a principle, the atom/bond tree is chosen according to the chemical structure of the molecule unless the structure would violate the requirement of a tree. So if the molecule has a covalent ring, then a bond is chosen to be omitted from the atom/bond tree. In a whole protein chain, the chain of backbone nitrogen and carbon atoms is the main-chain, and those atoms take precedence over other atoms [21]. Within the side chain, the non-hydrogen atoms closer to the main-chain take precedence; so C^{α} takes precedence over C^{β}, and C^{β} over C^{γ}, and so on. A chain of C^{α}, C^{β}, and the

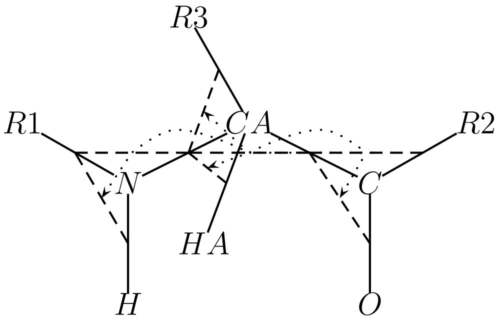


FIGURE 28. Z-System of "Backbone"

atoms that follow is considered as the main-chain of the side chain. If two carbons are bonded to the carbon in a main-chain, then the one with the lower numbering takes precedence over the other. For example, if $C^{\gamma 1}$ and $C^{\gamma 2}$ are bonded to C^{β}, $C^{\gamma 1}$ takes precedence over $C^{\gamma 2}$. The bond/angle tree and the angle/wedge tree have been chosen so that the edges of the tree would be growing out of the edges along the main-chain.

4.1.1. Backbone.

The (partially labeled) Z-system Γ^* for "backbone" which can be created using the program IMIMOL [10] is shown in Figure 28. This Z-system of "backbone" will be glued with another "backbone" on each side and a side chain to construct the Z-system of a protein chain, so the Z-system has three fake atoms that are not in the chemical structure of the backbone. $\{R1\}$ and $\{R2\}$ will be destroyed in gluing backbones, and $\{R3\}$ will be used to glue on a side chain. The labels are assigned as if the side chain and backbones are glued already since gluing only adds a new tetrahedron but does not change existing labels. The bond length and bond angles are taken from the parameter file of the CHARMM force field, the collection of empirical force field topology and parameter files [20].

Because of the tendency to be in a low energy state, non-bonded atoms normally tend to be as far from each other as possible. Because of that nature, some wedge angles are commonly known to have certain values, and those angles are also assigned. For example, the improper wedge angle from the backbone triangle $\{N, CA, C\}$ to

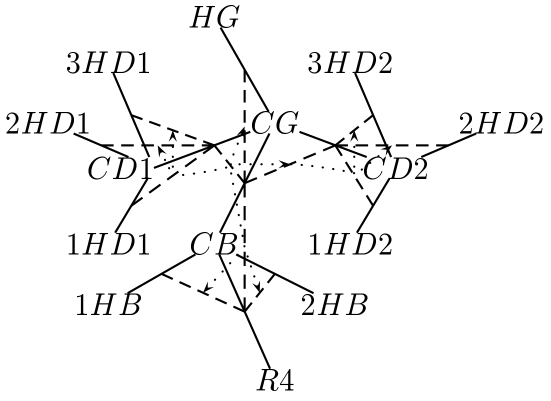


Figure 29. Z-System of the Leucine Side Chain

the side chain triangle $\{N, CA, R3\}$, which is $[N, CA, C, R3]$ with the side chain represented by $\{R3\}$ here, is always positive, about 120°, and the one to $\{N, CA, HA\}$, that is $[C, N, CA, HA]$, is negative, about $-120°$. This is an important and amazing feature of biological proteins since nonbiological synthesis yields a mixture of this type and its mirror image [22]. It is still a mystery as to why the wedge angle to the side chain is always positive in the biological proteins.

Often the two wedge angles along the backbone $[R1, N, CA, C]$ and $[N, CA, C, R2]$, or less precisely, the rotations about the bonds $\{N, CA\}$ and $\{CA, C\}$, are called ϕ and ψ wedge angles respectively. These angles vary and thus are not yet numerically labeled. This is why this Z-system is only a partially labeled Z-system.

4.1.2. Side Chains. A Z-system for each side chain is also constructed in IMIMOL [10]. An example, the Z-system of leucine, is shown in Figure 4.1.2. (For a complete catalog of Z-systems for the side chains, see the Appendix B.) Like the Z-system for backbone, each side chain also includes a fake atom $\{R4\}$ which would be used in gluing the side chain to the backbone. For the side chains that have rings in their chemical structures, one or more covalent bonds must be omitted to obtain atom/bond trees.

Name	χ_1	χ_2	χ_2 range
pp	62	80	
tp	-177	65	
tt	-172	145	120 to 180
mp	-85	65	45 to 105
mt	-65	175	

TABLE 3. Rotamer Table for Leucine [19]

Under normal biological conditions, some amino acid side chains have parts which are approximately rigid. A methyl group, -CH$_3$, is an example of such a part. Leucine has two methyl groups, one of which is composed of the atoms $CD1$, $1HD1$, $2HD1$, and $3HD1$. This methyl group is connected to the rest of the side chain via the bond $\{CG, CD1\}$ and is rotatable about the axis of this bond $\{CG, CD1\}$. So we choose the wedge $[CB, CG, CD1, 1HD1]$ to be the only dihedral and the improper wedges $[1HD1, CG, CD1, 2HD1]$ and $[1HD1, CG, CD1, 3HD1]$ are used to control the shape of the methyl group. In this way, we can control the orientation of the whole methyl group relative to the rest of the molecule via the single number labeling $[CB, CG, CD1, 1HD1]$. Other impropers in leucine and some wedges in the other side chains are chosen as impropers rather than dihedrals for similar reasons. Another example of a rigid body is a ring structure in a side chain, such as phenylalanine, in which the ring is essentially planar.

Excluding such rigid body structures and tetrahedra that involve hydrogen atoms, there are some free wedge angles. Such free wedge angles within the side chains are designated by χ_i where i represents the rotatable bond: $i = 1$ for the rotation about the bond $\{CA, CB\}$, $i = 2$ for the rotation about the bond $\{CB, CG\}$, and so on. Some study has been done on the χ-angles, and it was found that those angles fall into certain statistical patterns [19]. For example, leucine has two free wedge angles χ_1 and χ_2, and according to [19], those two angles falls into one of five patterns when leucine residues are studied in a large sample of high resolution protein structures. Table 3 lists some information taken from [19] where the name of each pattern, called

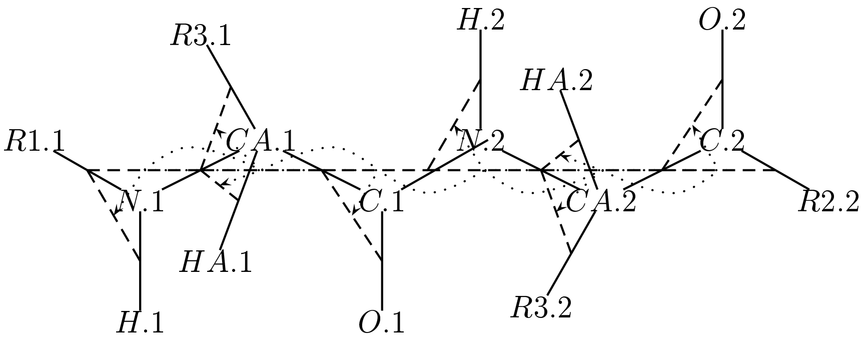


FIGURE 30. Z-system of Glued Backbone

a **rotamer**, and the common-atom value for each χ angle is given. (For a complete explanation, see [19].)

4.1.3. Protein Chain. To construct the Z-system for a chain of amino acids, we first construct a chain of "backbone" by gluing copies of the Z-system "backbone" shown in Figure 28. Assume we have a generic partially labeled Z-system (Γ^*, γ) for backbone on $\mathcal{N} = \{N, H, CA, HA, C, O, R1, R2, R3\}$. We let $\mathcal{N}.1 = \{N.1, H.1, CA.1, HA.1, C.1, O.1, R1.1, R2.1, R3.1\}$ and similarly for $\mathcal{N}.2$. Let $\Gamma^*.1$ and $\Gamma^*.2$ be Z-systems on $\mathcal{N}.1$ and $\mathcal{N}.2$, respectively, isomorphic to Γ^* on $\mathcal{N}$. Let $\gamma.1$ and $\gamma.2$ be partial labelings of $\Gamma^*.1$ and $\Gamma^*.2$ respectively derived from the partial labeling γ of Γ^*. The sites used in gluing are $(R2.1, C.1, CA.1)$ from $\Gamma^*.1$ and $(R1.2, N.2, CA.2)$ from $\Gamma^*.2$, and the new Z-system is $\Gamma.1 *_\mu \Gamma.2$ on $\mathcal{N}.1 *_\mu \mathcal{N}.2$ with $\mu = \{(R2.1, C.1, CA.1), (R1.2, N.2, CA.2)\}$, shown in Figure 4.1.3. In the process, the atom $\{R2.1\}$ is replaced by the atom $\{N.2\}$, and $\{R1.2\}$ by $\{C.1\}$. A new tetrahedron along the peptide bond $[CA.1, C.1, N.2, CA.2]$ is added during gluing to complete the angle/wedge tree. The wedge angle spanning the peptide bond is denoted by ω and is usually around 180°. Now we have a Z-system for a two-residue backbone. Gluing two copies of this, we obtain a Z-system for a four-residue backbone. Gluing two copies of this four-residue Z-system gives a Z-system for an eight-residue backbone. In this way, we are able to obtain a Z-system for a long chain rather quickly.

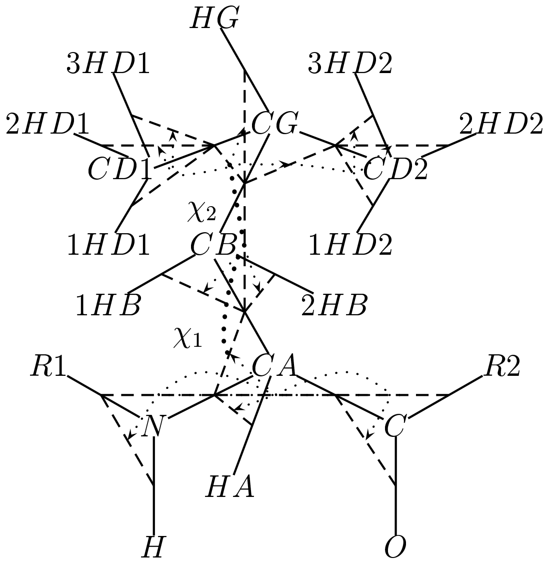


FIGURE 31. Z-System of Leucine Side Chain Glued to Backbone

Once we have the Z-system of the backbone chain, the Z-systems of the side chains need to be glued to it. Let Γ^* on $\mathcal{N}$ be the (partially labeled) Z-system of the backbone and Λ^* on $\mathcal{M}$ be the Z-system of the side chain. With leucine, for example, by gluing using the sites $(R3, CA, N)$ from Γ^* and $(R4, CB, CG)$ from Λ^*, we obtain a new Z-system $\Gamma *_\mu \Lambda$ on $\mathcal{N} *_\mu \mathcal{M}$ with $\mu = \{(R3, CA, N), (R4, CB, CG)\}$. Figure 4.1.3 shows the Z-system of leucine amino acid glued to the Z-system of the backbone. In this case $\chi_1 = [N, CA, CB, CG]$ and $\chi_2 = [CA, CB, CG, CD1]$ shown as dark dotted lines in Figure 4.1.3. The χ angles for a specific amino acid are indicated in the catalog in the Appendix B.

For a proline side chain, we glue on the side chain as described above and delete the $\{H\}$ attached to $\{N\}$ from the Z-system to agree with the chemical structure of proline on the backbone. It should be noted that we do not glue glycine to the backbone, but instead, the atom names HA and $R3$ on backbone would be changed to $1HA$ and $2HA$ respectively. Also the bond length and angle involving $\{2HA\}$ should be changed as necessary.

The atom names need to be adjusted as the side chains are glued so that in the end, each atom would have its atom name and amino acid name together with its residue number as its label. So if the leucine is the fifth amino acid in some protein sequence, then the C^{α} would have the label $CA.LEU5$. By gluing side chains in appropriate sequence, we have a Z-system of a protein chain. In this way, we construct the Z-systems of α- and β-apoproteins of *acid.* and a part of β-apoprotein of *moli.*

For the protein with the modified termini, the Z-system should also be modified to match with the chemical structure. The N terminus of the β-apoprotein from *acid.* is modified to have a carboxyl group, so the first residue (which is methionine) with the carboxyl group is named "CXM," or "X" as the amino acid name instead of "MET," or "M." As shown in the heirarchy tree of an apoprotein in Figure 23, the carboxyl group is the N terminus cap and can be also thought of separately from the first residue.

4.2. Z-systems of BCLs, Carotenoids, and Lipids

Since the chemical structure is the same for all three BCLs in the PC, it is enough to create one Z-system for a BCL. For carotenoids also, it suffices to have one Z-system if a PC only contains one kind of carotenoid as in the *acid.* structure given in the PDB file.

There are many ways to choose the trees of the Z-systems, especially for BCL. We use the same principle as in choosing the trees of the Z-system for the protein chain when applicable. The atom/bond trees are chosen according to their chemical structures for both BCL and carotenoid. However, since BCL and RG in *acid.* contain rings in their structures, we need to choose which bonds to omit. Other ways are possible, but for the Z-system of BCL, we have chosen in such a way that each tree looks like it is growing out of the magnesium atom in the head of BCL. In choosing the bond/angle tree, we need to be careful not to include any angle whose value is close to 0° or 180°. For the tail part of BCLs and carotenoids, the trees are constructed to grow out of the long chain of carbon atoms. The wedges for the angle/wedge trees are also chosen in a similar manner, and the orientations are assigned (the canonical orientation for the dihedrals). The Z-systems of these molecules and of a lipid are found in the Appendix B. The Z-system of a lipid is rather simple for it is mainly a long chain of carbon atoms with hydrogen atoms, and the head part attatched to the chains. We choose the carbon atom of the glycerol (to which the hydrophilic head part is bonded) to be the central atom and construct each tree going out of that atom and along the long chain of carbon atoms, and similarly for the head group.

4.3. A Z-system for a PC

To construct a Z-system for the PC, we tether the Z-systems of all the pieces: the α- and β-apoproteins, BCLs and carotenoids. For atom names to be distinct, to the existing labels, we add "A" for α and "B" for β for the Z-systems of the apoproteins, and "$BCLA1$," "$BCLA2$," and "$BCLB$" for the Z-systems of BCL B850a, B800, and B850b, respectively. The appropriate labels are added also to the carotenoide: "$RG1$" and "$RG2$" for the RGs in *acid.* and "$SPO1$" and "$SPO2$" for the spheroidenes in *sph.* These labels are separated by "." from the existing labels to indicate the different levels in the hierarchy in the organization of the Z-system.

Actually, it makes more sense that these pieces should be separate components in a GZ-system (see Chapter 1). However, to determine the $\mathcal{A}$ matrices labeling the linkages between components, we would typically employ an auxiliary Z-system. The tethered Z-system we describe here can be considered as having this auxiliary character.

In choosing the sites of tethering, we should be careful that a pair of sites chosen will not cause the angles created by tethering to be close to 0° or 180° to avoid the singular point of Z-system coordinates. Also, all sites should be, if possible, chosen where the structure of the molecule is rigid. A small change in the structure at the site will cause a big change in the relative position of the pieces being tethered. So we should also consider the variation present among the different PC structures in the PDB file, as discussed in Section 4.4.1. A good tethering site has small variation among different PCs, indicating rigidity.

4.3.1. Apoproteins. When tethering the Z-systems of the α- and β-apoproteins, sites should be chosen where the sequences are reasonably well-conserved between the species. The sites for tethering of the Z-systems of the two apoproteins are (CA, N, C) of T39.A.*sph.* and of P47.B.*sph.* since these residues are located in the middle of

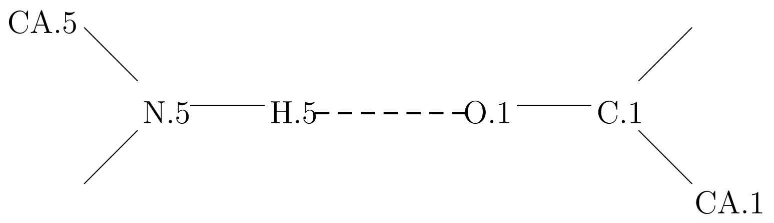


FIGURE 32. Selection of Sites

one of the most conserved regions comparing the sequences of *acid.* and of *sph* (see Figure 27).

4.3.2. BCL. To tether a Z-system of BCL or of carotenoid to the Z-system of the heterodimer, natural choices of the tethering sites are places of their ligations or hydrogen bonds. We first need to check if the ligation residue or the residue of the hydrogen bond is conserved between the species. If it is conserved, the sites should be chosen carefully since the atoms involved in hydrogen bonding and atoms bonded to those hydrogen-bonded atoms are typically almost linear, creating a bond angle of close to 180°. For example, in the backbone, if $O.1$ bonded to $C.1$ has a hydrogen bond with $N.5$ bonded to $H.5$ as shown in Figure 32, those four atoms are likely to be linear; thus, we should not choose $(O.1, C.1, CA.1)$ nor $(H.5, N.5, CA.5)$, but could use $(C.1, CA.1, O.1)$ and $(N.5, CA.5, H.5)$.

The contacts that hold the head parts of BCLs are the hydrogen bonds between magnesium atoms and the peptide atoms. The magnesium atom of B850a, BCLA1, is ligated to the nitrogen atom, NE2.H31.A.*sph.*, and the B850b, BCLB, magnesium atom is ligated to NE2.H39.B.*sph.* Observing the multi-sequence alignment in [14], the histidines that hold the B850 BCLs are highly conserved. Thus we assume that BCLs are ligated to those histidines in general. Since the magnesium atoms of B850 BCLs are ligated to the NE2 of the conserved histidine residues on the peptides, we choose the site $(NE2, CD2, CG)$ of the histidine residue on each α-apoprotein Z-system for BCLA1, and of the β-apoprotein Z-system for BCLB, for each to be tethered to the site $(MG, N\text{-}A, N\text{-}B)$ of BCLA1 or BCLB.

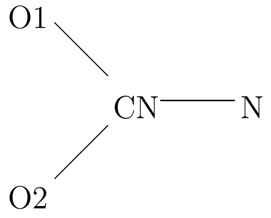


Figure 33. Carboxyl on N Terminus

One of the oxygen atom of the carboxyl group, O1.X1.A ligates the magnesium atom on B800. Since LH-II of *acid.* shares strong sequence homology at the N terminus of the α-chain with most of the LH-II complexes from non-sulphur purple bacteria [23], we assume that *sph.* α-apoprotein also contains the carboxyl group on the N terminus. Thus, $(CN, N, O1)$ of the carboxyl group on the N terminus would be the site to be tethered to $(MG, N\text{-}A, N\text{-}B)$ on the Z-system of B800, BCLA2.

4.3.3. Carotenoid. For carotenoids, we also need to make sure that the site and its structure are preserved despite the change of the types of the carotenoids between the species. One of the carotenoids in *acid.*, RG1, has two hydrogen bonds between conserved residues and atoms in its head: O2 with OE.E10.B.*acid.* and O6 with NZ.K5.A.*acid.* [23]. Knowing that RGs are used as templates for the spheroidenes, we compare the structure of RG and of spheroidene as in Figure 25 before deciding on the site. By comparing the two, we see that the ring part of RG is replaced by hydrogen atoms, thus O2 and O6 will disappear in the change that will be made. Also at the beginning of the tail, RG has all single bonds whereas spheroidene has conjugated bonds. Considering these structural differences, we choose the site $(C4, C5, C6)$, which is preserved between species and is close to O2, to be tethered to (CA, N, C) of Y14.B.*acid.* Although Y14.B.*acid.* is not conserved between the two species (the corresponding residue is Q23.B.*sph.*), we accept it as the site of tethering since it is on the protein backbone, the structure of which is carried over.

For the other carotenoid, RG2, no information is given in [23] on hydrogen bonding with the peptides. So we choose the site (CB, CA, N) of a conserved residue A36.B.*acid.* (A45.B.*sph.*), which is relatively close to the site $(C4, C5, C6)$ on RG2.

4.3.4. The Reference Atoms. We also tether the reference atoms $\{x_0\}$ and $\{x_1\}$. These two fake atoms serve as the coordinator of the PCs in building the Z-system of the whole LH-II since the ring structure consists of identical PCs. The atom $\{x_0\}$ serves as an origin for the ring of LH-II. The tethering site on the reference atoms is a 1-site (x_0, x_1) where the bond $\{x_0, x_1\}$ is along the rotational symmetry axis of the PCs.

Observing from the xyz coordinates of the atoms given in 1NKZ.pdb, the origin $(0, 0, 0)$ is closer to the periplasm side; thus, we choose the site (CA, N, C) of Q46.A*sph.* and tether the reference atoms. If we were to tether to the site closer to the cytoplasm side, we would get one of the new angles closer to 0°.

4.4. A Labeled Z-system

To label a Z-system, IMIMOL [10] has the capability to calculate the labels of the Z-system if the configuration is given in a .xyz file format. Since an atomic structure file in the PDB is most often given in a .pdb file format, which also gives the configuration of a system of molecules by providing xyz coordinates for each atom, we need to convert it to a .xyz file format. First, however, since most PDB files do not include the hydrogen atom positions, we add those positions to the .pdb file by using the program MolProbity [7], which uses all-atom contacts and geometrical criteria to add the hydrogens. Then, we convert this .pdb file with hydrogens to a .xyz file format. Babel is a widely used program to convert one file format to the other, but since atom names are not converted correctly in the conversion from a PDB file to a XYZ file, we use the program Convert, written by Jason Rogers, which converts the configuration part of a .pdb file to a .xyz file format. Then by "imposing" a .xyz file on the matching Z-system using IMIMOL, we obtain the labeled Z-system.

4.4.1. PDB. Files in the PDB contain much useful information beside a configuration and amino acid sequences. For example, some files provide information on ligation sites or secondary structures of the molecule. In the .pdb files of LH-IIs from *acid.* and from *moli.*, we also find that certain matrices are given. This is because the configuration given in the files is not for the complete ring of LH-II but only for a few PCs. The PDB file for *acid.* gives a configuration for three PCs, and for *moli.* for four PCs. The given matrices are active translation $\mathbf{b}$ and rotation matrices A, and by applying each of them to the given configuration R, one may obtain the configuration for the full ring structure of LH-II.

However, the PCs given in the PDB file have not only different configurations but different conformations. In other words those PCs are not identical, so the LH-II structure obtained from applying the matrices to the given configuration will not result in a fully nine-fold symmetric LH-II.

The differences in the PCs are the result of the way multiple LH-II rings pack together to form a crystal studied by X-rays to obtain the atomic level structure. This variety, however, gives us useful information about the flexible and rigid regions in the LH-II complex when it is subjected to packing stresses and held at a temperature of 100 K (body temperature is about 300 K and 1° C= 1 K).

4.4.2. Averaging PCs. To build a nine-fold symmetric LH-II, we need one *acid.* PC which we can use as a structural base to build a *sph.* PC. Since there are three different PCs of *acid.* and we do not know which conformation is closer to the structure of *sph.*, we decided to average those conformations. So an averaging program was written to average the different numerical labels of the identical Z-systems. This program writes a Z-system with averaged labels and also calculates the standard deviation of each label. The averaging program takes different labeling of the same Z-system and averages the different labels for each edge. The average and the standard deviation of bond lengths and bond angles are taken in the normal way, but since wedge angles have a periodic nature, from $-180°$ to $180°$, a different method is used. Let $\theta_1, \theta_2, \ldots, \theta_n$ be the wedge angle labels of a particular wedge where n is the number of different configurations for the PC. Then define $r > 0$ and $\bar{\theta}$ to be such that

$$re^{i\bar{\theta}} = \frac{e^{i\theta_1} + e^{i\theta_2} + \cdots + e^{i\theta_n}}{n}.$$

We take $\bar{\theta}$ as the "average angle" of $\theta_1, \theta_2, \ldots, \theta_n$. When r is close to 1, the standard deviation is small, and when r is close to 0, we understand the standard deviation to be large [11].

In order to use this program, a .pdb file with multiple PCs is converted to a .xyz file format, and the .xyz file is separated into different .xyz files each containing the configuration of one PC. Since we will use the averaged labels as a starting value for the tethering labels of the reference atoms discussed in Section 4.3.4 and Z-system of a PC, we add the coordinates of the two reference atoms to each .xyz file. The matrices given in the PDB file for *acid.* tells us that the axis of rotation of PCs is along the z-axis with no translation; thus, we assign the coordinates $(0, 0, 0)$ and $(0, 0, 1)$ to the reference atoms $\{x_0\}$ and $\{x_1\}$ respectively.

By imposing .xyz files for three different PC conformations to the same Z-system of the reference atoms and a PC, we obtain three different labeling of this Z-systems. In the process, IMIMOL [10] warns that the coordinates of atom $H24.RG2$ in the Z-systems is not specified in the file for any of the three xyz files. Although MolProbity [7] did not add the hydrogen atom to $C24.RG2$, since the chemical formulae given in the PDB file are the same for both RG's in the PC, we decided to keep the hydrogen atom $H24.RG2$ and assign the labels by hand for the edges involving the atom. The bond $\{C24, H24\}$, the bond angle $\{C23, C24, H24\}$ and the wedge angle $[C25, C23, C24, H24]$ are labeled 1.1 angstrom, 120.873° and 180°, respectively, by averaging the labels for the same bond, angle and tetrahedron of the RG1s of the three PCs and rounding up.

We also learn that in one of the PCs, the coordinates of $HE2.H41.A$ is not specified and the atom $HD1.H41.A$ in the xyz file is not in the Z-system. So H41.A has a hydrogen on NE2 and not on ND1 in the Z-system whereas a hydrogen is on ND1 and not on NE2 for this PC according to the .xyz file. Since IMIMOL did not give a warning for the other PCs, it means that MolProbity [7] has assigned a different protonation state for this histidine residue in this particular PC. Since the other two PCs have the same protonation state, we follow the protonation state of the majority. So we assign the labels for edges involving $HE2.H41.A$ in the minority PC by assigning the average of the labels of the other two PCs.

Now since we have three labeling of the same Z-system, we run them through the averaging program and obtain the averaged PC. Then we check to make sure that the tethering angles are not close to 0° or 180° and also check the standard deviations around the tethering sites to measure the rigidity.

4.4.3. Proline. Once we obtain an averaged Z-system, we need to adjust the conformation of proline since its ring is not rigid. A non-rigid ring has to satisfy certain constraints to stay as a ring with a reasonable bond lengths and angles, and the average conformation probably will not satisfy those constraints. In such a case, the bond which is present chemically but has been omitted from the tree of the Z-system has an unreasonable length which is not the averaged bond length from the three different PCs. So to preserve the average bond length for the bond $\{N, CD\}$, for a proline ring, we used the formula derived in [9]. So let l_{01}, l_{12}, l_{23}, l_{34}, and l_{04} be the bond lengths of the bonds $\{N, CA\}$, $\{CA, CB\}$, $\{CB, CG\}$, $\{CG, CD\}$, and $\{N, CD\}$, respectively. Let θ_1, θ_2, and θ_3 be the bond angles of $\{N, CA, CB\}$, $\{CA, CB, CG\}$, and $\{CB, CG, CD\}$, respectively. Also let χ_1 and χ_2 be the wedge angles for $[N, CA, CB, CG]$ and $[CA, CB, CG, CD]$, respectively. Then

$$\chi_2 = 2 \arctan \left(\frac{B \pm \sqrt{A^2 + B^2 - C^2}}{A + C} \right)$$

where

$$\begin{aligned}
A &= 2 l_{34} \sin\theta_3 [l_{01}(\cos\theta_1 \sin\theta_2 + \sin\theta_1 \cos\theta_2 \cos\chi_1) - l_{12} \sin\theta_2] \\
B &= 2 l_{34} \sin\theta_3 l_{01} \sin\theta_1 \sin\chi_1 \\
C &= l_{04}^2 - (l_{01}^2 + l_{12}^2 - 2 l_{01} l_{12} \cos\theta_1) - (l_{23}^2 + l_{34}^2 - 2 l_{23} l_{34} \cos\theta_3) \\
&\quad -2[l_{01}(\cos\theta_1 \cos\theta_2 - \sin\theta_1 \sin\theta_2 \cos\chi_1) - l_{12} \cos\theta_2](l_{23} - l_{34} \cos\theta_3).
\end{aligned}$$

Using the averaged values for l_{01}, l_{12}, l_{23}, l_{34}, l_{04}, θ_1, θ_2, θ_3, and χ_1, we get two values for χ_2; so we choose the one closer to the averaged χ_2 given in the Z-system of the averaged PC.

4.5. PC for *sph.*

Following the decision made as discussed in Section 3.3.2, we use *moli.* as our template for the beginning of the *sph.* β-apoprotein. We prepare the averaged labeled Z-system of *moli.* β-apoprotein and glue residues 3 to 13 (the PDB file for this species is missing information on the first two residues) onto the Z-system of *acid.* β-apoprotein residues 9 to 41 so that in terms of *sph.*'s residue number, residues 7 to 17 are taken from *moli.* and residues 18 to 50 are from *acid.* The wedge angle added as a result of gluing is assigned the average of corresponding wedges from the two species. So the wedge angle $[CA.A17, C.A17, N.E18, CA.E18]$ is $-179.536°$ which is the average of wedge angles $[CA.S8, C.S8, N.E9, CA.E9]$ of *moli.* and $[CA.A13, C.A13, N.I14, CA.I14]$ of *acid.*

For the apoprotein parts of the Z-system of the averaged PC with the patched β-apoprotein, we change the sequences of amino acids to that of *sph.* We also change the carotenoids to match with the chemical structure of spheroidenes, or SPOs.

Usually building a model by means of homology modeling requires highly technological tools to simulate chemical forces and other environmental factors so that a final product will reflect a natural low free energy state for that particular sequence of amino acids. However, we are more interested in building a model by focusing on the geometrical aspects of molecules, so we will use the exact backbone structure of the templates, without using a simulation tool. We will keep the exact backbone structure from the template, at least for the major part, and we use the rotamers as much as possible for the side chains since those are the common conformations of a particular side chain. By using homology modeling and working by hand, we hope to see and have a good grasp of what kind of changes can be made without changing the major structure, such as the backbone structure, and what kind of changes affect the structure to what degree.

By changing the side chains to match with the sequences of the target, we expect to see some clashes, the overlap of the **van der Waals spheres** centered at two non-bonded atoms. We will try to resolve those clashes by adjusting minor parts of the structure. Our priorities in keeping the structure or the rotamers are, from the highest to the lowest: the backbone, conserved residues, then nonconserved residues. Among the nonconserved residues, the ones in more conserved regions have higher priority than the ones in less conserved regions. We work with the lower priority items to resolve clashes, and we adjust the higher priority items only when it is necessary.

For the conserved residues, since we do not believe that the exact conformation given in the averaged PC of the templates would carry over between the species, we assign the rotamer closest to the pattern of χ angles observed in the conformation of the averaged PC. However, we keep the conformations from the averaged PC of the templates for the conserved residues that have the tethering sites on their side chains since changing the χ angles will change the position of the site, which in turn affects the relative positions of the two pieces tethered using that site. For the nonconserved residues, we also wish to choose a rotamer which will not cause a clash between two non-bonded atoms.

4.6. Conformation Assignment

To assign the rotamers to the nonconserved residues and to check for clashes, we need to be able to observe the structure closely. So we make the Z-systems of each apoprotein from the PC changed to the structure of *sph.* and view each structure using RasMol [28]. When a clash is detected, we change the labels of the wedge angles in that area using IMIMOL, then we check if the clash has been resolved or not using RasMol. If the clash is not resolved, we adjust the wedge angles using IMIMOl again till we resolved the problem. The complex adjustments described in Section 4.8 were done in a similar way but using VMD-IMI and the PowerWall.

4.6.1. Apoproteins-Conserved Residues. First, we check the rotamerized conserved residues to see if there is any clash caused by rotamerizing the side chains. When a rotamerized conserved residue clashes with the backbone atoms, since the backbone has higher priority to be kept in the given structure than the conserved residue to be in its rotamerized conformation, we use the averaged conformation for that residue. In some cases, the clash is already present in the averaged PC. Then we check the individual PCs to see if the clash is caused by averaging. T46.B had clashes with the backbone atoms, and the clashes were present in the averaged PC as well as an individual PC. We tried to resolve or soften the clashes, but it seemed that T46.B would clash with some surrounding atoms in any position, so it was left as in the averaged PC structure. Some rotamerized residues such as V30.B also clashed with a backbone atom, but the clashes were resolved simply by turning the methyl group (that is, by changing the dihedral wedge angle which control the orientation of the methyl group), so those residues are still in their rotamers.

If two conserved residues clash due to rotamerization, then we try to resolved the problem by adjusting the χ angles within the ranges given in the rotamer library [19]. Some atoms of W40.A were too close to the surrounding atoms in either

Residue	Average	Rotamer	Other
K5.A	-58, -161, 180, -142	-67, 180, 180, 180	-82, 180, -140, -150
W7.A	-49, 111	-65, 95	-49, 100
H31.A	-80, 83	-65, 80	
Y44.A	175, 54	-177, 80	
E16.B	-71, 171, -43	-67, 180, -10	
E18.B	-68, -80, 0	-65, -65, -40	-68, -80, -20
E19.B	-74, 177, -1	-67, 180, -10	
H21.B	-151, 60	-177, 60	
R29.B	-76, 164, 178, 177	-67, 180, 180, 180	-76, 180, 180, 180
F31.B	-166, 99	-177, 80	
L36.B	-77, 170	-65, 175	-77, 175
H38.B	-77, 79	-65, 80	
T46.B	-105	-65	
W48.B	174, 55	-177, 90	

Table 4. Non-Rotameric Conserved Residues

the rotamerized or the averaged conformation, so we chose what appeared to be the least conflict with the surroundings which was to assign the rotamer angle for χ_1 and the angle within the given range for χ_2. If adjusting within the ranges does not resolve the clash, then we use the conformation observed in the averaged PC. H21.B and I25.B clashed into each other by rotamerizing both. We sought to resolve the clash by assigning the angle within the range and turning the methyl group of I25.B, but later when the structures of α- and β-apoproteins were put together, we found that H21.B had a hydrogen bond with X1.A. To preserve the hydrogen bond, H21.B was assigned the averaged conformation, and that resolved the clash with I25.B. The final conformation assigned to each conserved residue with a non-rotameric conformation is shown in Table 4. The χ angles given in the “Rotamer” column are from the rotamers closest to the average conformation. The residues with no entry in the column of “Other” have the averaged conformations, and the ones for which some angles given in the “Other” column have those angles as their conformations.

4.6.2. Apoproteins-Nonconserved Residues. After adjusting the conserved residues, starting from the residues in more conserved regions and moving toward less

Residue	χ angles	Residue	χ angles
T2.A	62	V34.A	-60
N3.A	-65,, -20	T36.A	-65
L8.A	-65, **-100**	T37.A	62
K11.A	-67, 180, 180, 180	L41.A	-172, 145
T13.A	-65	Y45.A	**-55, -40**
V14.A	-60	S48.A	-65
V16.A	-60	V51.A	175
L18.A	**-155, 86**	V7.B	175
F19.A	**-137, 86**	W8.B	-65, 95
S21.A	**75**	V14.B	**73**
V24.A	**-55**	V20.B	-60
I25.A	-65, 170	Q23.B	-65, -65, -40
S27.A	-65	L24.B	-85, **40**
V28.A	-60	L26.B	-172, 145
I29.A	-65, 170	M34.B	-67, 180, 75
I30.A	-65, 170	I37.B	-65, 170

TABLE 5. Nonconserved Residues

conserved regions, we choose the rotamers for the nonconserved residues. For each residue, we try different rotamers from the rotamer library for that amino acid till we find one which gives a clash-free conformation; then, we do the same for the next residue in a lesser conserved region. If no rotamer gives a clash-free conformation, we go back to the previous residue to see if another rotamer would give a clash-free conformation for that residue so that the next residue may have a rotamer with a clash-free conformation. When that fails, then we look for another conformation which is clash-free. For some residues, the first rotamer we tried fit well, and for some other residues we had to try all possible rotamers. Yet for some others, none of the rotamers gave a reasonable conformation so that we needed to make adjustments, and these are discussed in more detail in Section 4.8. The final conformation for each nonconserved residue to which a rotamer could have been assigned is shown in Table 5. The angles that are non-rotameric are shown in bold.

4.6.3. Heterodimer. After adjusting and assigning the conformation for all the residues for each apoprotein, we tether the Z-systems of two apoproteins and check for the clashes between the two apoproteins in the similar manner. T46.B clashed also with conserved W40.A badly, but it was so in the averaged as well as in an individual PC structure. So as previously decided, it was kept in its averaged conformation. The conserved residues E18.B and K5.A clashed where they did not do so in the averaged structure. The averaged conformation for K5.A and the rotamer for E18.B did not resolved the clash, but the averaged conformation for E18.B and the rotamer for K5.A did. So E18.B was put to its averaged conformation, and K5.A was kept in its rotamer (although, the conformation of K5.A was changed later which did not affect the decision made for E18.B).

4.6.4. BCLs and SPOs. Then to the Z-system of the heterodimer, we tether the Z-systems of BCLs and SPOs one at a time, checking for the clashes between those and the heterodimer. The flat head part of BCLA1 clashed with conserved F31.B and since there was no clash in the averaged structure, F31.B was put to its average conformation. BCLA1's tail clashed with several other nonconserved residues as well. Two of them, S21.A and V24.A, were adjusted from their rotamers to the conformations in which the atoms appeared to have least contact with the surrounding atoms. Other residues required more extensive adjustment and are discussed in Section 4.8.

BCLB and BCLA2 seemed to be more removed from the apoproteins and did not require a major adjustment. In one place where a clash was detected, it was resolved by changing the orientation of the methyl group of the tail of one of the BCLs.

The clashes observed when the Z-system of SPO1 was tethered to the Z-system of the heterodimer and the BCLs were all with nonconserved residues. Q23.B was assigned another rotamer which resolved the clash. I29.A clashed with the tail part of SPO1 toward the end where it was a double bond in RG but is a single bond in SPO, which means the part becomes flexible. So the wedge $[C25, C26, C27, C28]$ was changed to 105° from about 90° in RG. With some methyl groups' orientation adjustments, the clash was resolved without changing the conformation of I29.A from its rotamer. Another residue, L24.B, was in the only rotamer possible within the β-apoprotein, so we searched and assigned a non-rotameric conformation so that the clash with SPO1 would be resolved. The residue F19.A was originally in the rotamer conformation which occupied the similar space as its corresponding residue L19.A in *acid.*, but it created a problem with a methyl group in the middle of the tail of SPO1. Since turning the methyl group did not resolve the problem, we assigned another rotamer which resolved the clash with SPO1. However, later we found that that conformation caused clashes with the neighboring PC atoms, described in Section 4.8 also. SPO2 clashes with one of the BCLs which is from the averaged structure, but it did not cause many clashes with the side chains since the tail wanders out into space. One conserved, rotamerized residue Y44.A did clash with SPO2, so that residue was put to its average conformation.

4.6.5. Two PCs. After adjusting the conformation within the PC, we put two copies of that same PC structure together to adjust the conflicts between the PCs. (How we tethered the two copies of the Z-system of the PC is explained in Section 4.7.) The only clash for which the adjustment was simple was the clash between L18.A of one PC with the tail of BCLA1 of another PC. To avoid the crowdedness between the PCs, L18.A was moved to the non-rotameric conformation facing the inner space of the ring of LH-II. The other clashes were chain-reaction clashes which required more work, and how we attempted to resolved them is discussed in Section 4.8.

4.7. A Z-System of LH-II

Once we have resolved the clashes between two PCs, we have the PC structure which can be used to form a ring of LH-II made up from nine identical PCs. Since we coordinate lipid molecules relative to a PC, we tether one on each of the cytoplasm side and the periplasm side of the β-apoprotein part of the Z-system in such a way that the lipids are on the outside of the LH-II ring. The sites are (CA, N, C) of E18.B and F40.B on the peptide and $(C1, C2, C3)$ on the Z-system of lipid. We name the lipid in our structure "PEL" for phosphatidylethanolamine with lenoleic and lenolenic acids as the tails. So we add "$PEL1$" for the Z-system of a lipid on the cytoplasm side and "$PEL2$" for the lipid on the periplasm side.

To build the Z-system for the LH-II, we tether eight PCs to the Z-system of the reference atoms and a PC. The site on each PC is the same as the first PC tethered, which is the site (CA, N, C) of Q46.A. For the site on the reference atoms, we use $(x_0, x_1, CA.Q46.A)$ where $CA.Q46.A$ is the atom on the PC previously tethered. So to tether the second PC, we add "$.PC1$" to the atom labels of the first PC and "$.PC2$" to the one which is going to be tethered. The sites used to tether are $(x_0, x_1, CA.Q46.A.PC1)$ and (CA, N, C) of Q46.A.PC2. The operation is as described in Section 1.6, and the tethering labels are the same as for the first PC with

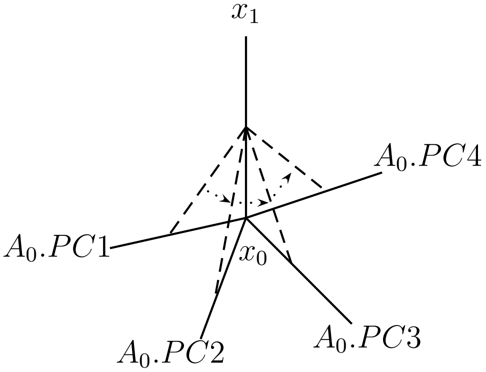


Figure 34. Tethering of the Reference Atoms and PCs

$[CA.Q46.A.PC2, x_0, x_1, CA.Q46.A.PC1]$ being 40°, which is the rotation angle between the two PCs. The next PC is labeled "$.PC3$," and the site of the reference atoms is $(x_0, x_1, CA.Q46.A.PC2)$, the reference atoms and the atom of the PC previously tethered. In this way, the rotation angles between the PCs are the same, creating a symmetric ring.

4.8. Complex Conformational Adjustment

4.8.1. V51.A. The nonconserved residue V51.A, whose corresponding residue in *acid.* is lysine, looked as if it was smashed with the backbone. In the template structure, the lysine does not clash with the backbone since it is long but not as wide as valine. The C terminus end of the α-apoprotein is sticking out to the periplaslm side and shows much variation between PCs in the template structure. Since we do not know the biological function of this part, sticking out to the periplasm, we took the liberty to change the backbone structure and adjusted several wedge angles around V51.A. We also changed along the backbone of A50.A since H.A50.A and H.V51.A on the backbone clashed in the template structure.

4.8.2. L8.A. The cytoplasm side where the α- and β-apoproteins come toward one another is very crowded and tightly packed. It is rather amazing that we did not see any worse clashes than what we observed. Probably, the worst clash in that region was the clash between L8.A and A17.B. (Recall that up to A17, the template is of *moli.* for β-apoprotein.) The hydrogens from those residues were so close that they could form a bond between them. We did not align the α-apoprotein sequence from *moli.* in Figure 27, but according to the multi-sequence alignment in [14], both residues are conserved from *moli.* So we tried the conformation similar to that leucine from *moli.*, but it did not resolve the clash. Since the structure of the α-apoprotein is from *acid.* and the structure of the β-apoprotein in this region is from *moli.* which is patched to the structure from *acid.*, the relative position of the two residues were different from that in *moli.* So we attempted to move L8.A to another conformation to resolve the clash. However, the cytoplasm side of L8.A was covered by another nonconserved residue, W8.B, which made it impossible for L8.A to move away from A17.B. W8.B was positioned in such a way that any move of W8.B using χ angles would cause even worse clashes. Since the N terminus of the β-apoprotein is probably sticking out into the cytoplasm side, we assumed that that part of the β-apoprotein structure would be flexible (although the standard deviation among the different PCs given in the PDB file of this region for *moli.* is small). So we moved the backbone wedge angle $[N.P9.B, CA.P9.B, C.P9.B, N.S10.B]$ by 10° so that W8.B would move away from L8.A, making room for L8.A to move away from A17.B. Since this move caused L8.A to approach V14.B, nonconserved V14.B was moved out of its rotamer toward the open space.

4.8.3. K5.A. The conjugated bonds starts closer to the head part in SPO than in RG so that the part which is flexible in RG is fixed flat in SPO. This change in the structure caused a clash between SPO1 of one PC and the conserved K5.A of neighboring PC. Although K5.A is conserved, the lysine residue is a flexible amino acid, so it is more reasonable to change the conformation of K5.A than to bend the flat part of SPO1. So we adjusted the conformation of K5.A to avoid the clash, but this region is still tightly packed, and probably not the exact structure in *sph.*

4.8.4. F19.A. As mentioned earlier, nonconserved residue F19.A clashed with atom in the middle of the SPO1 tail within the PC, and when it was moved to another rotamer, it clashed with the BCLA1 tail of the neighboring PC. So we moved it back toward the SPO1, keeping enough distance so that F19.A would not clash with SPO1. Since it still clashed with the end of the tail of BCLA1, the tail was moved away from the residue. However, since the tail would clash with some other atoms if we moved it away too much, we did not resolve the F19.A clash with the tail of BCLA1. This residue is in between the two apoproteins where the SPO1 slides in and between two PCs and where the tail of BCLA1 curls around the α-apoprotein. Moving one way or the other would cause a clash with some atoms of the pigments; therefore, we did not find the conformation which would resolve all the clashes in this region.

Chapter 5

Evaluation

5.1. Summary

5.1.1. α-apoprotein. We have 54 residues for α-apoprotein from *sph.* and have the structure for 53 residues from the template, *acid.*: 24 conserved and 29 nonconserved. Among 24 conserved residues, 13 residues can be rotamerized. (Glycine and alanine do not have a rotamer, and prolines are assigned their own conformations as described in Section 4.4.3.) Out of those 13 residues, nine are in their rotamers closest to the observed conformations in the averaged PC. The four residues that are not in their rotamers are K5, W7, H31, and Y44. Since H31 has the tethering site on its side chain, it is in its average conformation to avoid the change in relative position of the apoprotein and BCLA1. Y44 is also in its average to preserve the hydrogen bond with W48.B of the neighboring PC. A tryptophan is a big amino acid, and a part of W7.A is surrounded by the conserved residues. Since this residue in the template structure also has some clashes, W7 is assigned a conformation which appears to soften those clashes the most. Another residue, K5 is in a non-rotameric conformation, avoiding the clash with SPO1 as discussed in Section 4.8.3.

apoprot.	cons.	poss.	rotamer	noncons.	poss.	rotamer
α	24/54	13	9	29/54	23	18
β	29/50	20	10	15/50	9	7

Table 6. Summary of Rotamerizability

There are 29 nonconserved residues with 23 residues possible to be rotamerized. Eighteen residues are in their rotamers, and Y45 is in the conformation which creates a hydrogen bond with BCLA1. The other four residues, L8, L18, F19, and S21, are in non-rotameric conformation, avoiding clashes; although, F19 has some clash still present as discussed before.

5.1.2. β-apoprotein. There are 50 residues in the β-apoprotein from *sph.* for which no structure is available for the residues 1 through 6, the structure of residues 7 to 17 is from *moli.*, and that of 18 to 50 is from *acid.* As we can see from Table 6, 29 residues are conserved from the template species with 20 residues possible to be rotamerized. Ten out of those 20 residues are in their rotamers. Eight of non-rotameric conserved residues are in their average: H39 to keep the position of the tethering site with BCLB, H21 and W48 to preserve the hydrogen bonds, and E16, E18, E19, and F31 to avoid clashes. One more residue among those eight is T46, in which the severe clash is present even in the template structure; thus, this clash is not resolved. The other two non-rotameric conserved residues, R29 and L36, are assigned the average angle for χ_1 to avoid clashes and the rotamer angles for the other χ angles.

Nine out of 15 nonconserved residues are possible to be rotamerized, and seven of them are in their rotamers. χ_1 of L24 is in a rotamer with χ_2 five degree off from the range given for that rotamer. Another non-rotameric, nonconserved residue V14 is assigned the conformation which turns the residue away from the crowded region.

	acid.			*sph.*	
	Atom	Atom		Atom	Atom
	O1.X1.A	NE2.H12.B		O1.X1.A	NE2.H21.B
	SD.X1.A	NH2.R20.B		SD.X1.A	NH2.R29.B
	O.G4.A	OG.S8.B			
	N.W7.A	OG.S8.B			
	NE1.W7.A	ND1.H12.B		NE1.W7.A	ND1.H21.B
*	O.W7.A	N.L3.B		O.W7.A	OG.S10.B
	OD1.N11.A	N.A1.B			
	N.T39.A	OE1.Q46.A(+)		N.T39.A	OE1.Q46.A(+)
	OG1.T39.A	NE2.Q46.A(+)		OG1.T39.A	NE2.Q46.A(+)
	NE1.W40.A	O.W45.A(+)		NE1.W40.A	O.Y45.A(+)
	OH.Y44.A	NE1.W39.B(-)		OH.Y44.A	NE1.W48.B(-)
	NZ.K50.A	ND1.H41.B(-)			
	Mg.BCLA1	NE2.H31.A		Mg.BCLA1	NE2.H31.A
	OBB.BCLA1	NE1.W45.A		OBB.BCLA1	OH.Y45.A
	Mg.BCLB	NE2.H30.B		Mg.BCLB	NE2.H38.B
	OBB.BCLB	OH.Y44.A(+)	*	OBB.BCLB	OH.Y44.A(+)
	Mg.BCLA2	O1.X1.A		Mg.BCLA2	O1.X1.A
	OBB.BCLA2	NE.R20.B		OBB.BCLA2	NE.R28.B
	OBB.BCLA2	NH2.R20.B		OBB.BCLA2	NH2.R28.B
	O2.RG1	OE1.E10.B			
	O6.RG1	NZ.K5.A			

Table 7. Inter-Pigment Hydrogen Bonds

5.2. Discussion

5.2.1. Hydrogen Bonds. Table 7 lists the hydrogen bonds present in the *acid.* structure and the corresponding hydrogen bonds in the structure of *sph.* In the table, (+) indicates that the residue is of the counterclockwise next neighbor PC viewing the LH-II ring from the periplasm, and (-) for the residues on the clockwise next neighbor PC.

The hydrogen bonds between conserved residues or between a conserved residue and a backbone atom are mostly preserved. The bond between O.W7.A and N.L3.B is between backbone atoms, but because of the template change of β-apoprotein to *moli.*, this bond is not preserved. However, O.W7.A appears to pick up the bond with OG.S10.B. Some hydrogen bonds are preserved even with the change of amino acid such as the bonds of OBB.BCLA1 to NE1.W45.A in *acid.* and to OH.Y45.A in *sph.*

A nonconserved residue K11.A in any of its rotamers does not take part in the hydrogen bond, but by changing the conformation, it may bond with OE2.E18.B. Also, another nonconserved residue N3.A may be involved in the hydrogen bonding with atoms of BCLA2.

5.2.2. Hydrophobicity. The residues 12 to 37 in the α-apoprotein and the residues 14 to 45 in the β-apoprotein are the transmembrane α-helices in *sph.* structure. The corresponding residues in *acid.* for the α-apoprotein are the same, and in the β-apoprotein in *acid.* are numbered from 5 to 36. Comparing the amino acid sequences between those two species in these helices regions, many nonconserved residues are changes within the group of hydrophobic amino acids. Even with a few residues involving the changes with the hydrophobic and non-hydrophobic amino acids, we see that the balance between hydrophobic and hydrophilic residues is maintained.

5.2.3. Crowded Regions. The structure of *sph.* LH-II as well as of *acid.* is crowded in general but is amazingly well-packed. Some of the areas, such as around T46.B as mentioned earlier and also W7.A, are crowded, and clashes are present in the template structure.

Due to the changes in the side chains, the structural change in the carotenoids, and patching of the backbone for the β-apoprotein, the model structure of *sph.* has crowded areas that are not in the structures of the templates. Some regions such as nonconserved V24.A which is surrounded by the tail of BCLA1 are made to fit in the available space by being assigned a non-rotameric χ angles. Some other regions that are discussed in Section 4.8 required more extensive and careful adjustments. Even then, we have not found satisfactory conformations for some of those regions, such as around F19.A.

Reviewing those crowded regions unique to the structure of *sph.*, especially the regions of K5.A and F19.A where the residues clashed with SPO1, led us to think of the possibility of moving SPO1 to resolve the problem. In the effort of resolving the clashes of those difficult regions, we found that it is not sufficient to move one coordinate at a time. We also came across to the need of being able to move the structure locally without affecting large portions of the structure.

5.2.4. Evaluation and Future Improvement. Overall, we have used the internal coordinate system to build a fully symmetric LH-II ring and were able to keep the structure of the templates and cope with the clashes by changing only the end parts of molecules; although, a definitive conclusion cannot be made since some clashes are still unresolved. As for the use of rotamers, it certainly served as a guide in determining the possible conformations for certain side chains. However, since rotamers are "averages," we should expect some deviation from the exact rotamers, especially for the conserved residues. Also for nonconserved residues, some may be able to form hydrogen bonds by having non-rotameric conformations. So the idea of assigning rotamer conformations to all side chains needs to be re-evaluated as to what extent we should impose that principle.

As we have mentioned, the capability to work locally without affecting the whole structure is a necessity. We need to be able to fix two points and work in between them without changing any other parts. To do so, we need to implement a way to systematically change multiple coordinates. Detecting clashes and knowing how bad they are is also necessary to evaluate the resulting structure more precisely. Also, we would need a faster computer or better algorithms as well as the capability to implement the GZ-system formalism to build any system bigger than LH-II.

5.3. Conclusion

Except for a few parts that we have already described, we achieved building the model of LH-II in *sph.* following the principles. Yet the structure that we have obtained still has many uncertainties and variances, and we would not venture to claim that we have found *the* structure of *sph.* LH-II. Using the exact backbone structures from the templates for the most part, we could see that the structure was valid even with so many side chain changes. Especially for the α-apoprotein, none of the residues from 21 to 30 is conserved between *sph.* and *acid.*, and yet the backbone structure and pigment arrangements are such that side chain changes could be made without changing the core structures. This shows that even with the fact that different species have different amino acid sequences, the structures are remarkably similar. That fact makes one wonder about the basic design in that it can accomodate such varieties, and ask what kind of design it has. Then, even with the structures being so similar between the species, there are some places that caused us to pay special attention. Those areas remind us that they are unique species, and each has a unique structure of its own.

When we improve the things mentioned in the previous section, we may be able to resolve the clashes. However, there are many other things such as biological and chemical factors that come into play in what the exact structure would be. There are still many uncertainties and unknowns to what those factors are. Even when we think we have come to know those uncertainties and unknowns, those would be just the tip of an iceburg, and there will arise many more questions and wonders. Even when we think we have obtained the structure, there will be still much more to be unveiled. And again, we will be amazed at the deep, great wisdom and knowledge revealed even in such a thing as LH-II in a tiny bacterium. How much more, then, will we see the unfathomable wisdom and knowledge revealed in our own make-ups?

Appendix A: The Structure of LH-II

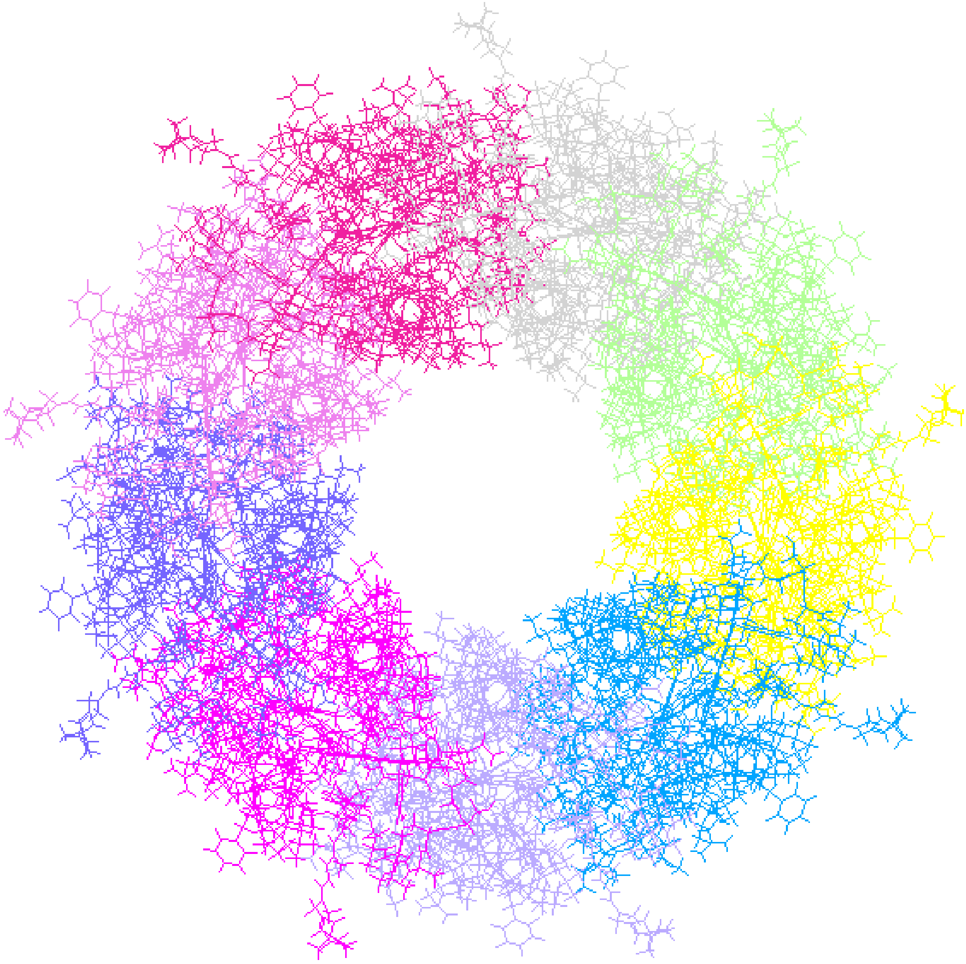


Figure 35. Nine PCs with identical conformations form a ring of LH-II.

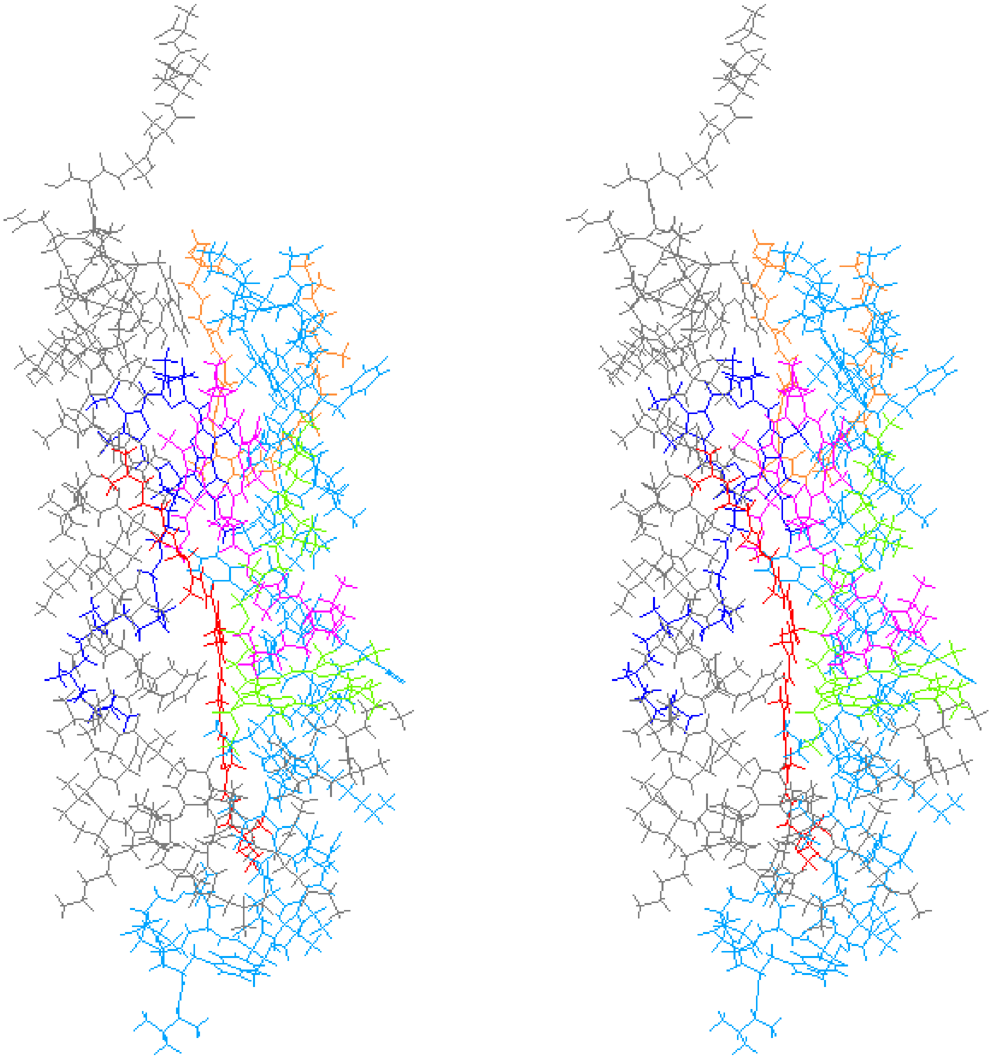


FIGURE 36. Cross-eye stereo view of the Structure of a PC. α-apoprotein in grey, β-apoprotein in skyblue, B850a in blue, B850b in magenta, B800 in green, SPO1 in red, and SPO2 in orange.

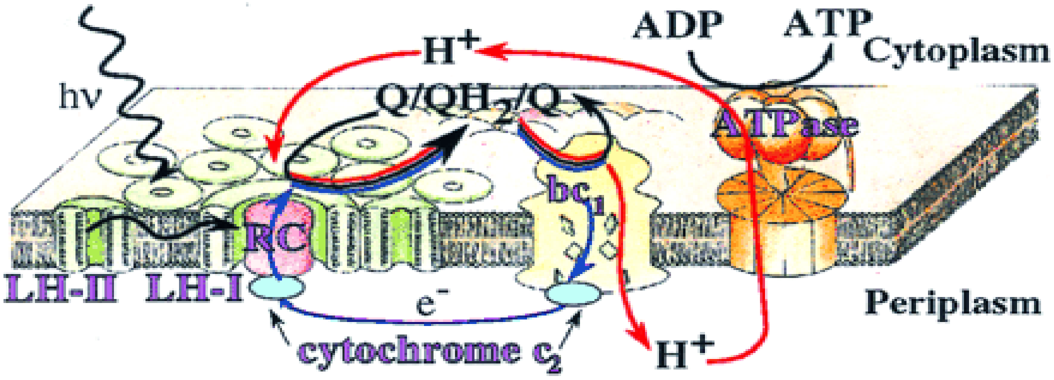


FIGURE 37. The Photosynthetic System in Cell Membrane

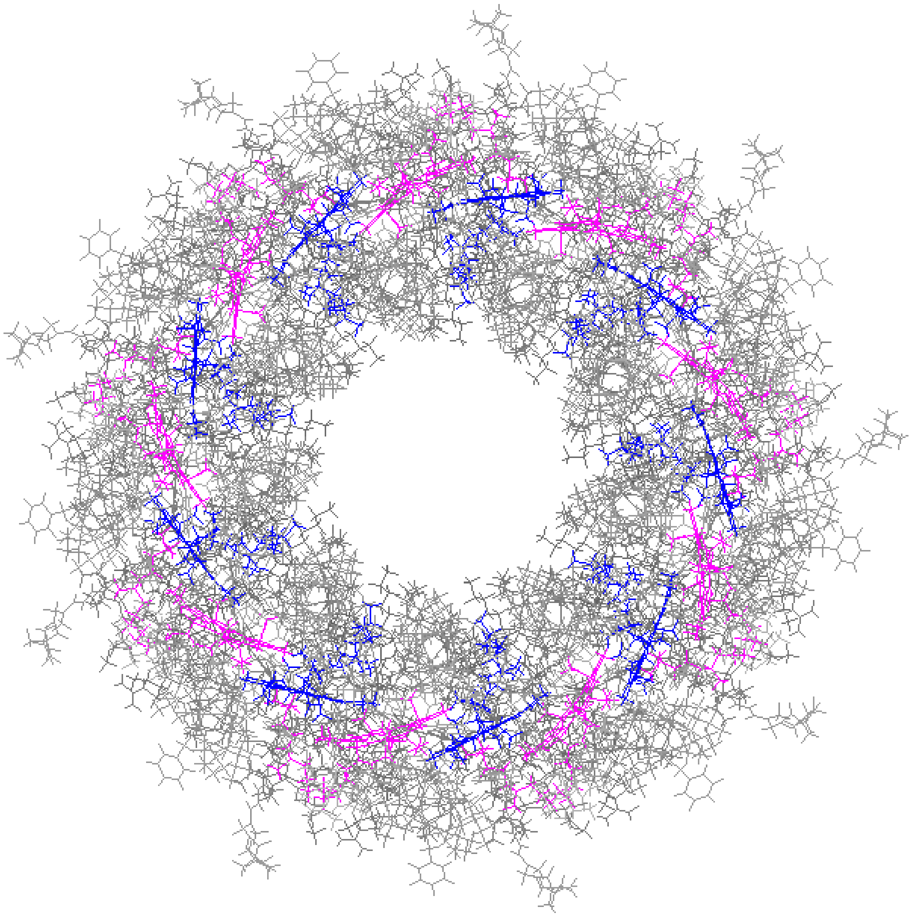


FIGURE 38. View from periplasm side. The B850a BCLs are in blue and B850b BCLs in magenta. The flat head parts of B850s (what look almost like lines from this view) form a ring, one overlapping with the BCLs next to it.

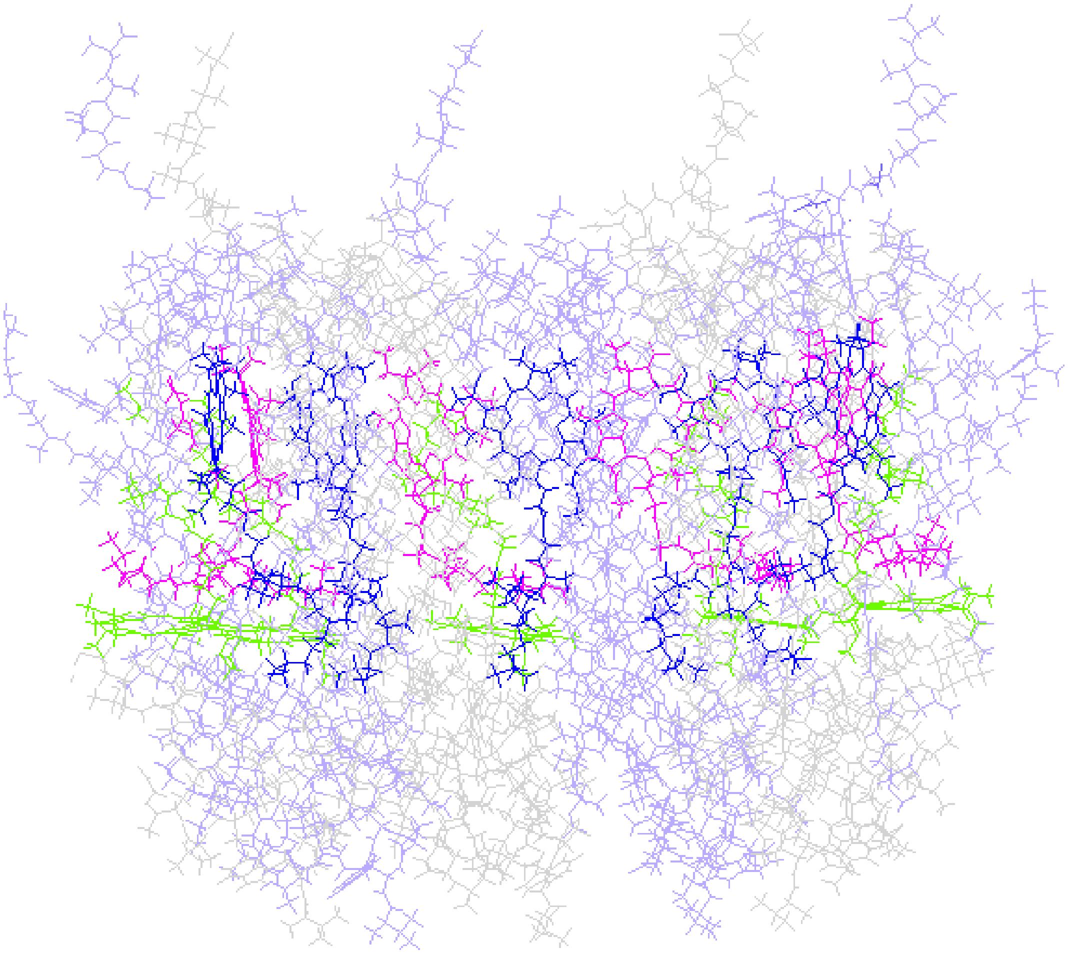
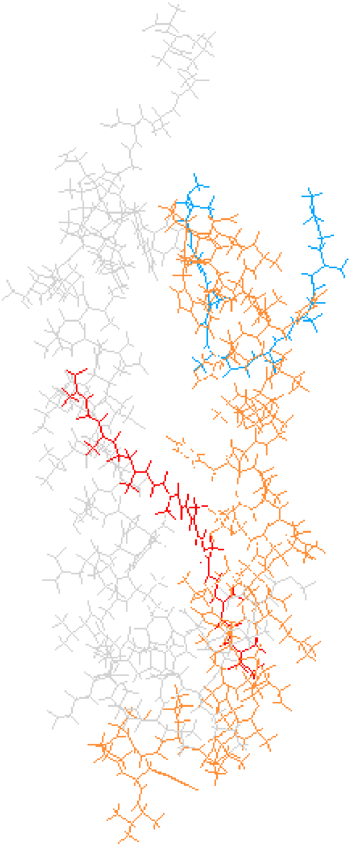


FIGURE 39. The view of LH-II from a cutting plane which is perpendicular to the membrane plane. Top is the periplasm side, and the bottom is the cytoplasm side. The B800 BCLs are colored with green (B850a with blue and B850b with magenta). We can see the flat heads of B800 lying under the tail of B850s, almost parallel to the membrane plane.



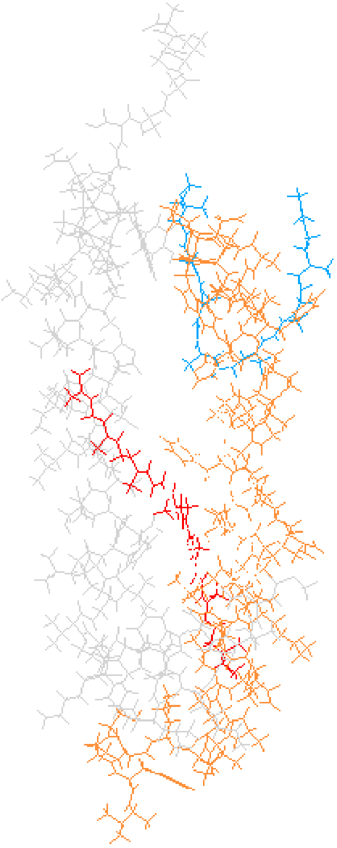


Figure 40. Cross-eye stereo view. SPO1 (red) runs through the two α-helices of the hetrodimer. The left side is the inside of LH-II ring.

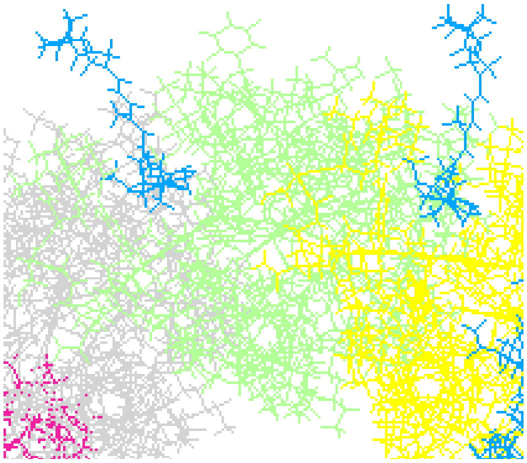


Figure 41. SPO2 (skyblue) occupies the space between two adjacent PCs.

Appendix B: The Z-systems

The followings are the Z-systems of amino acid side chains and pigments included in the structure of LH-II from *sph*. The atoms are labeled with their names; bonds are shown in blue; bond angles are in red; and wedge angles are in green. The ϕ, ψ, and χ wedge angles are also indicated.

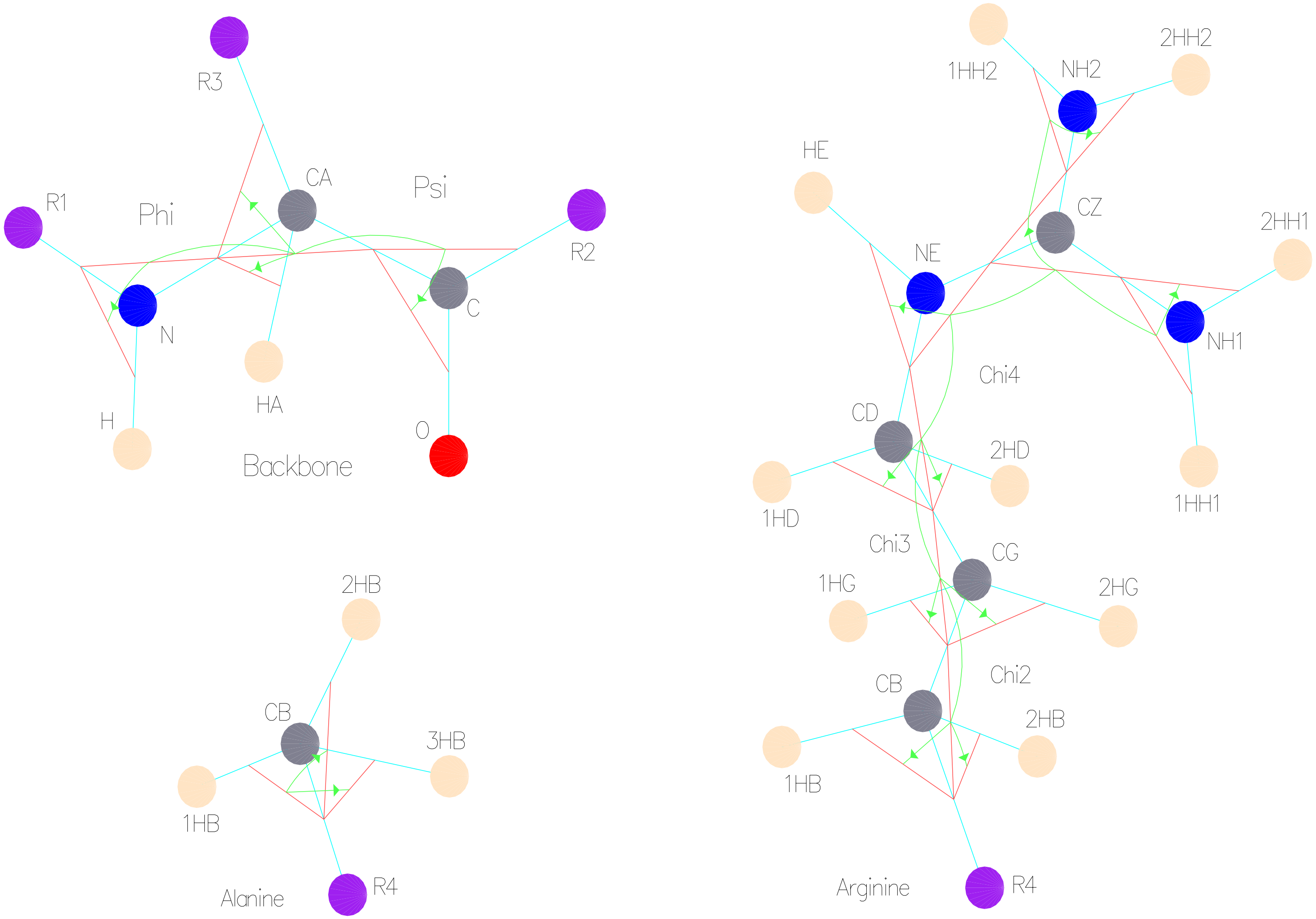


Figure 42. Backbone, Alanine, and Arginine

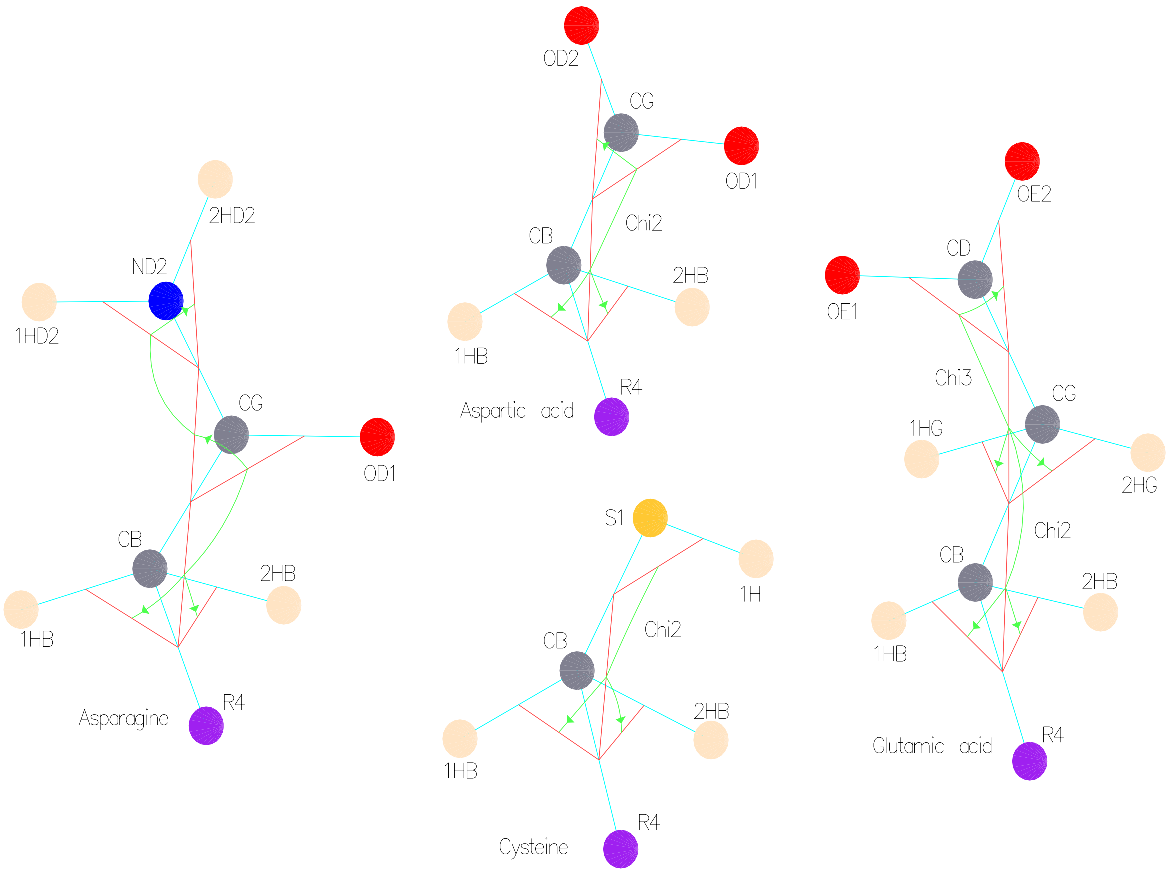


Figure 43. Asparagine, Aspartic Acid, Cysteine, and Glutamic Acid

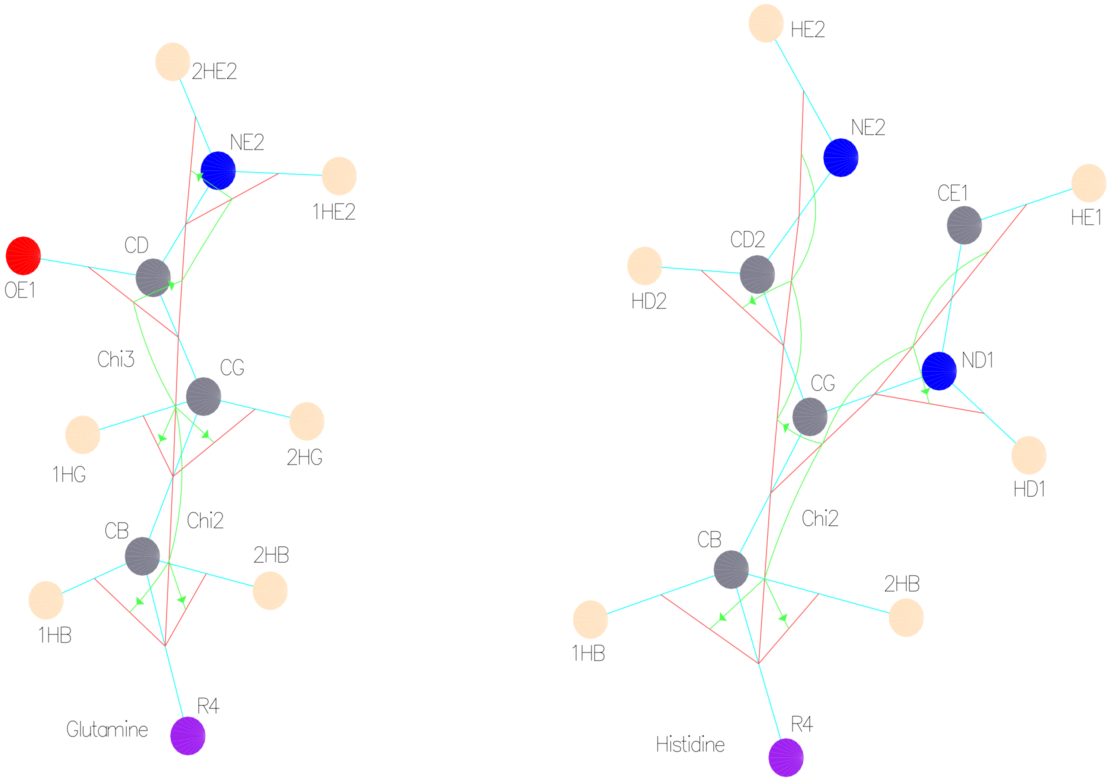


FIGURE 44. Glutamine and Histidine

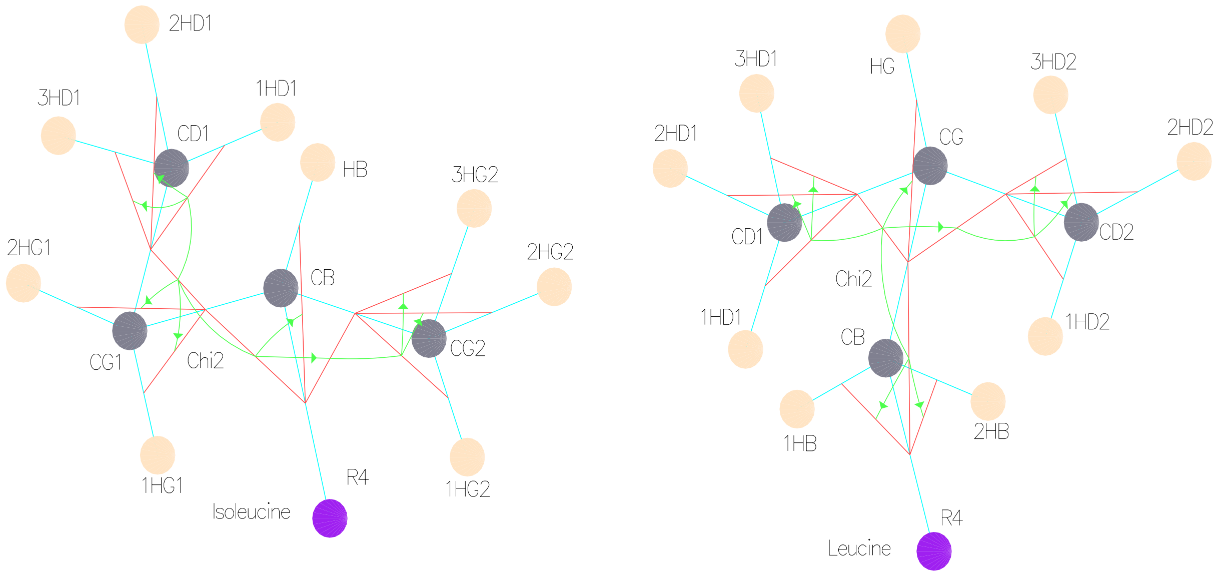


FIGURE 45. Isoleucine and Leucine

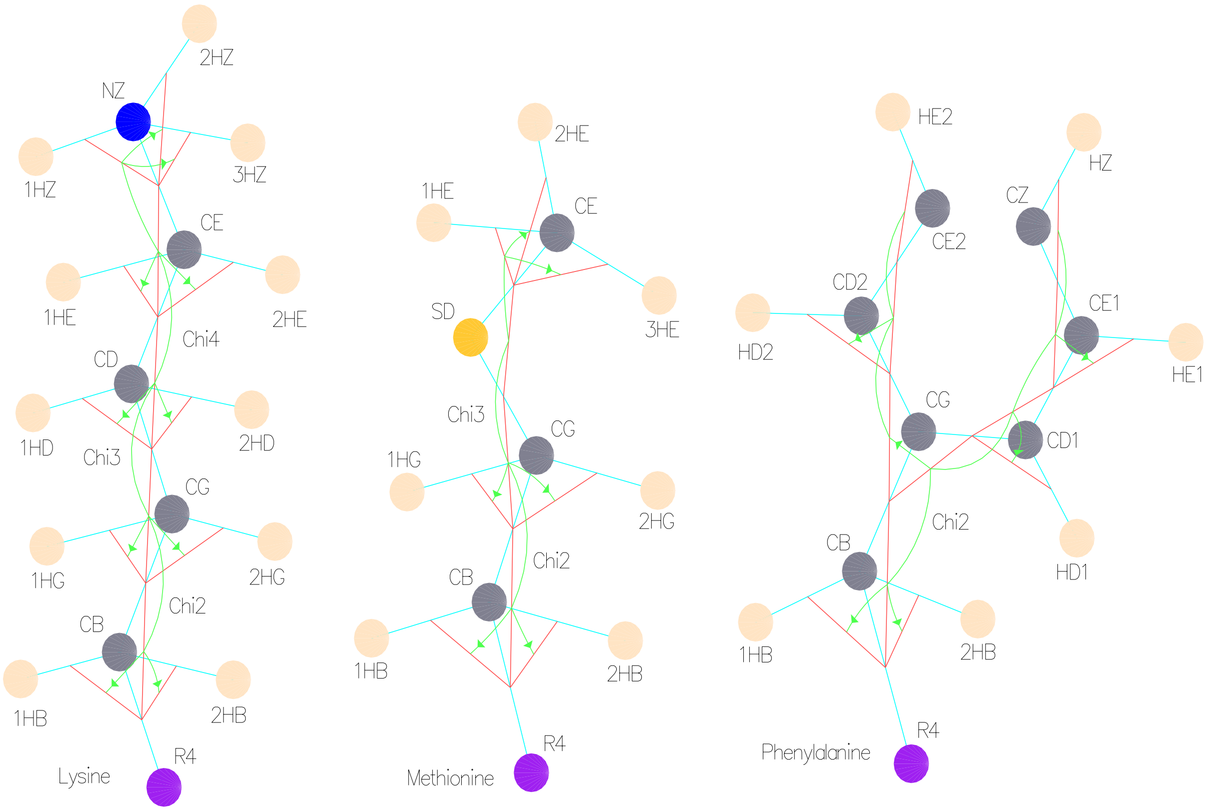


Figure 46. Lysine, Methionine, and Phenylalanine

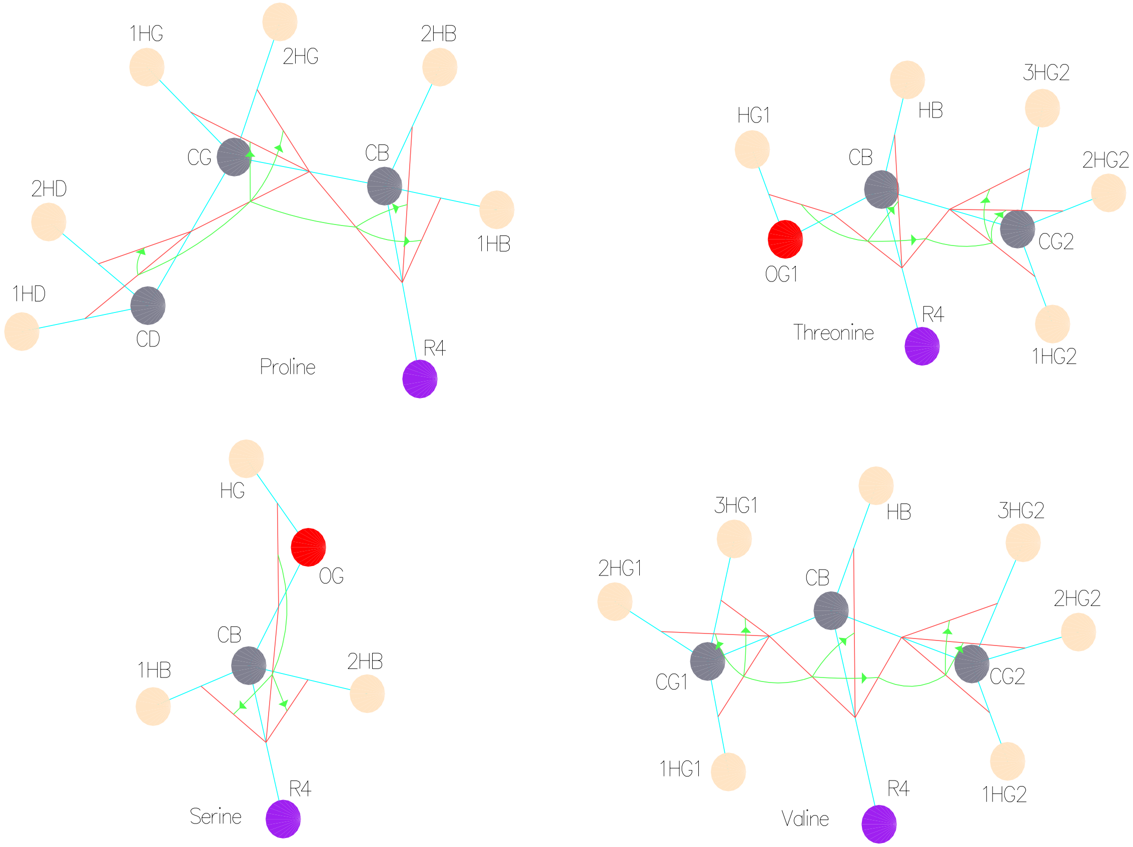


FIGURE 47. Proline, Serine, Threonine, and Valine

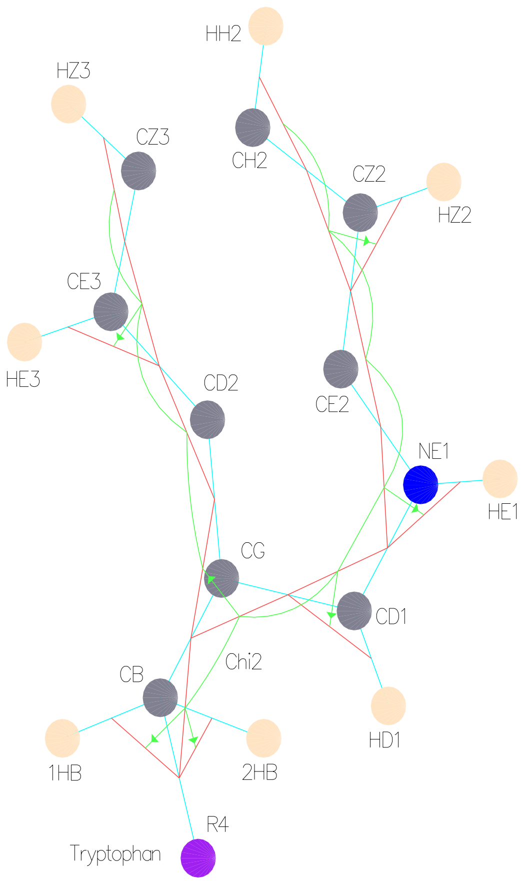
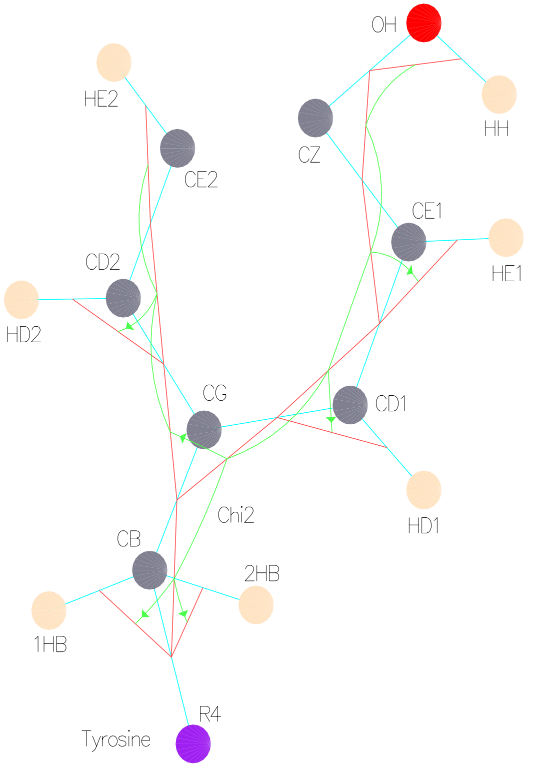


FIGURE 48. Tryptophan and Tyrosine



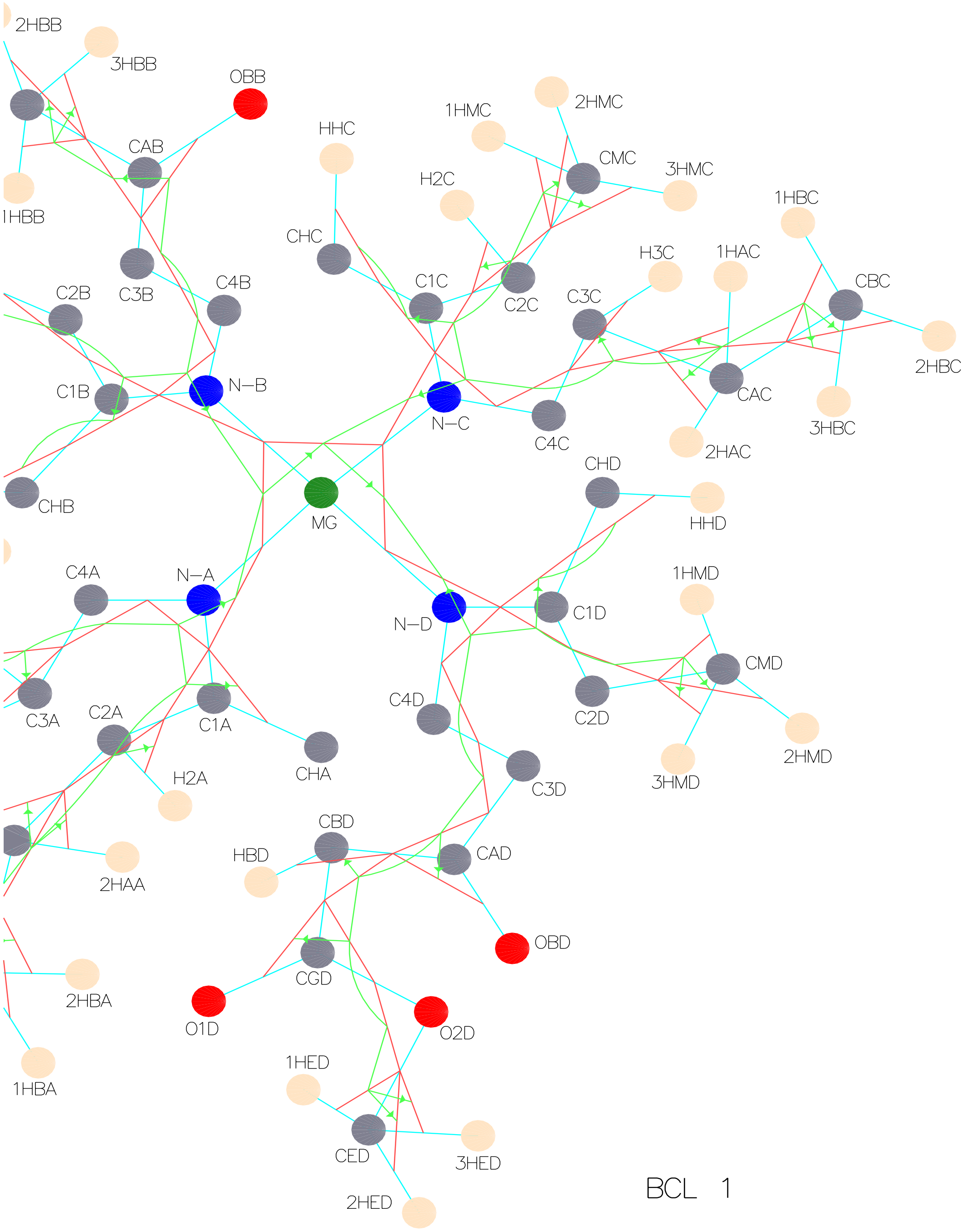


FIGURE 49. BCL 1

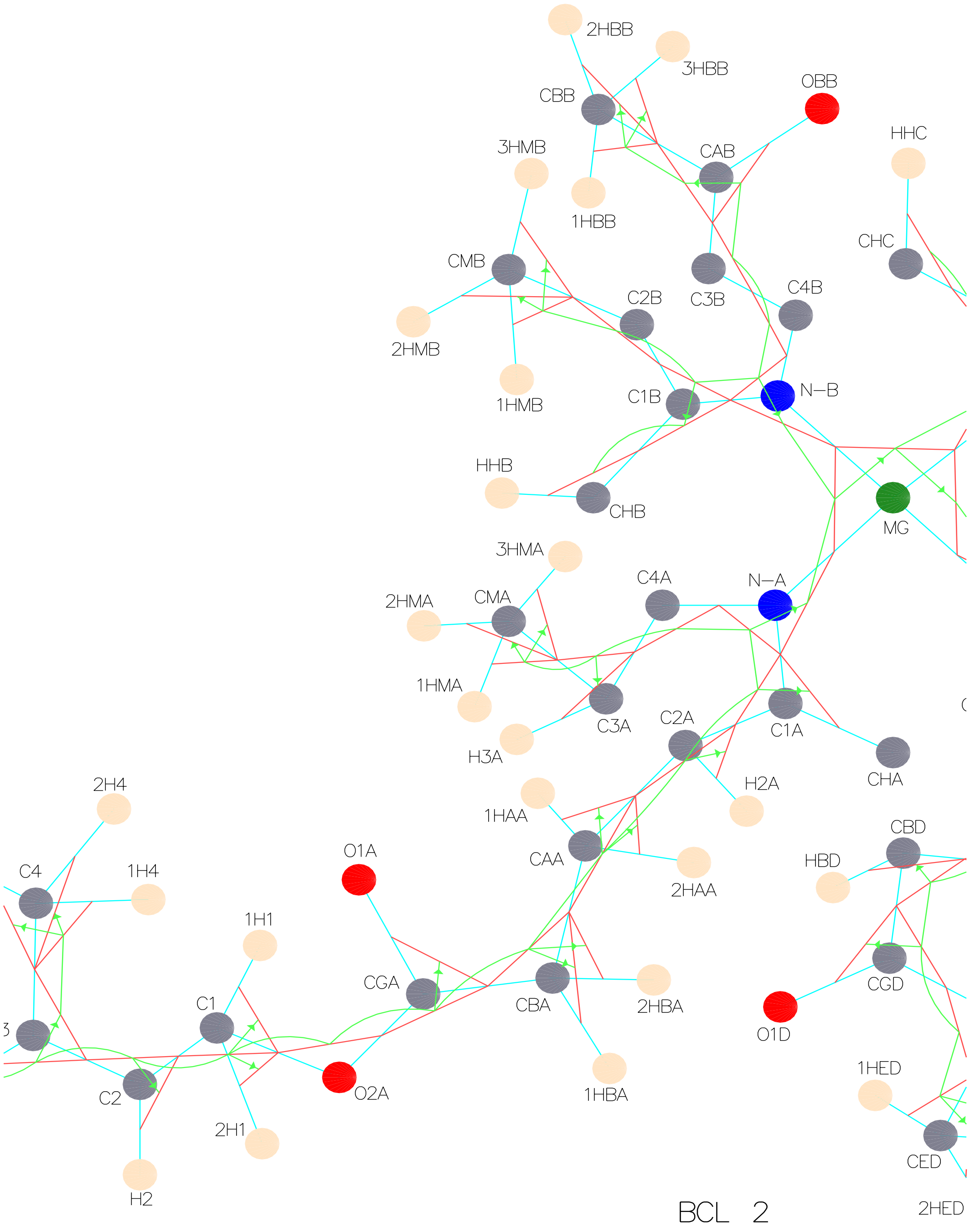


Figure 50. BCL 2

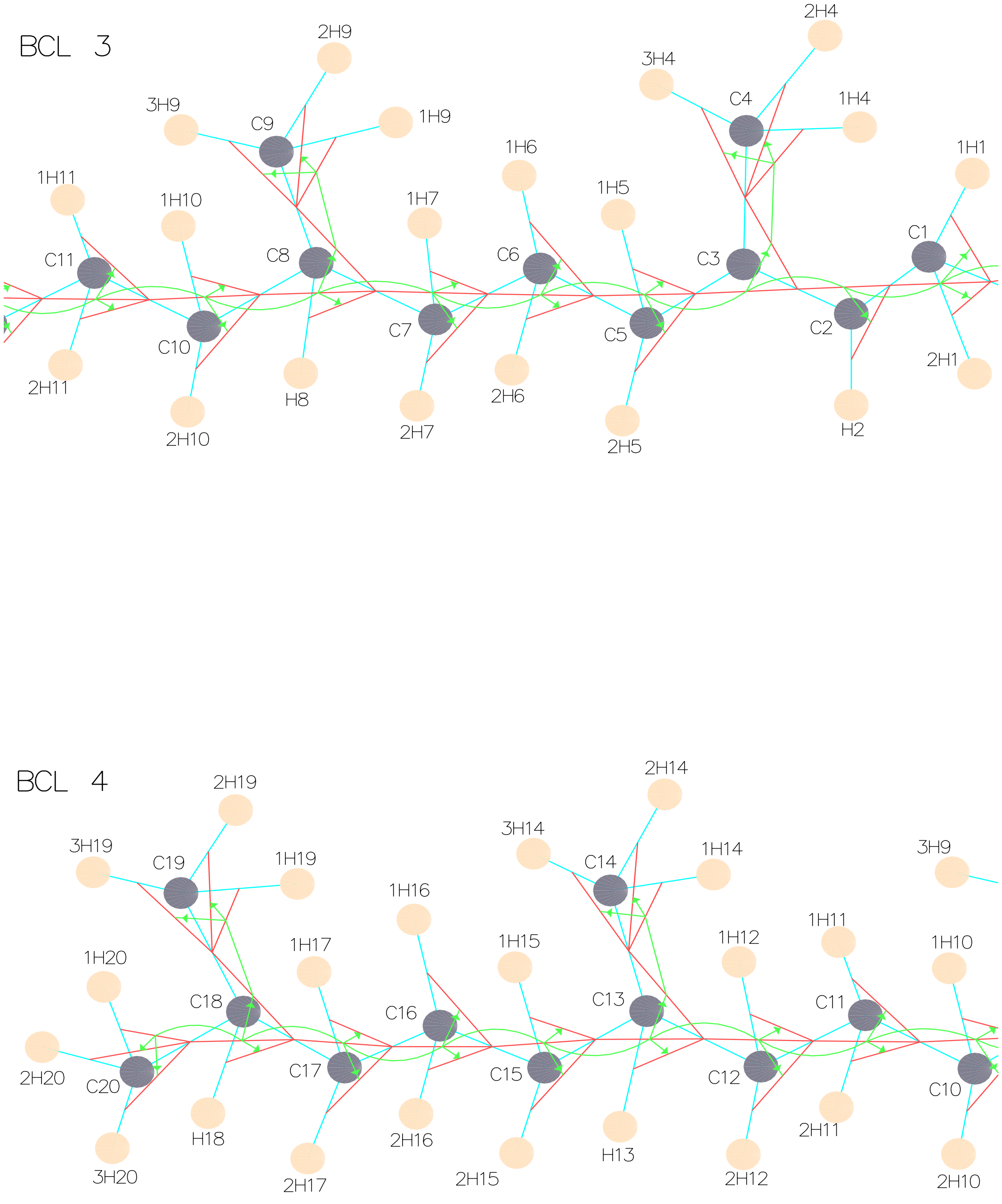


Figure 51. BCL 3 and BCL 4

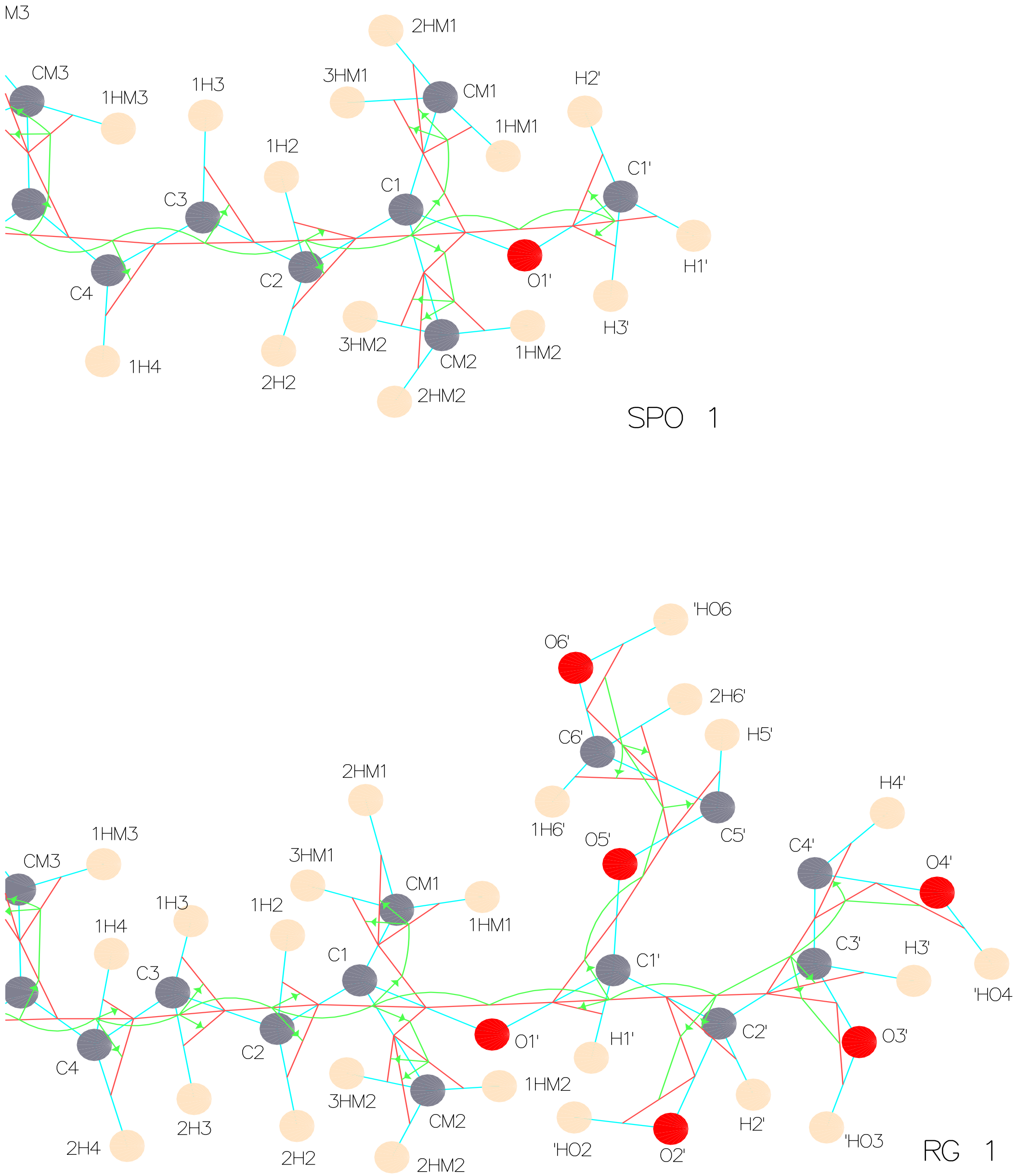


Figure 52. SPO 1 and RG 1

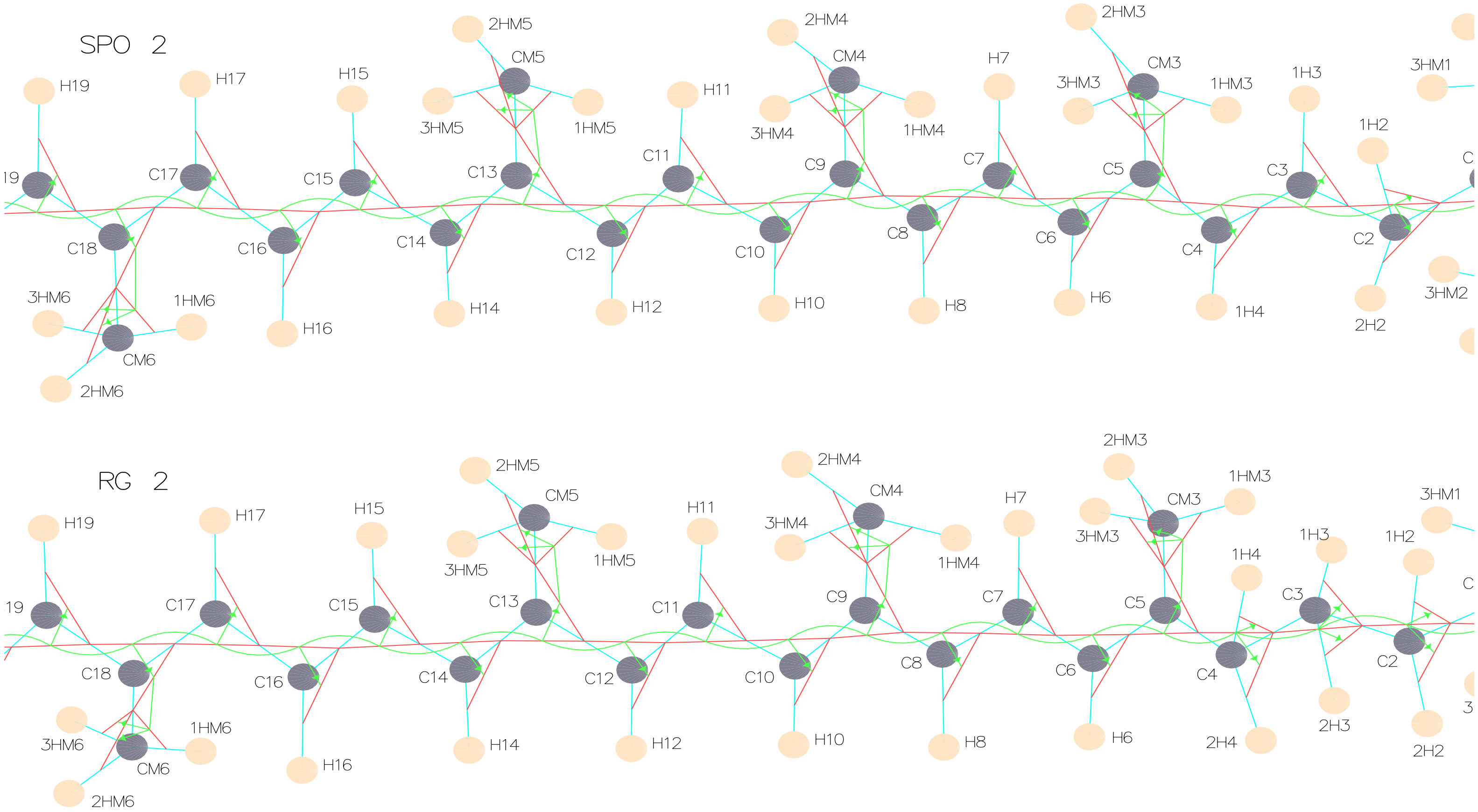


FIGURE 53. SPO 2 and RG 2

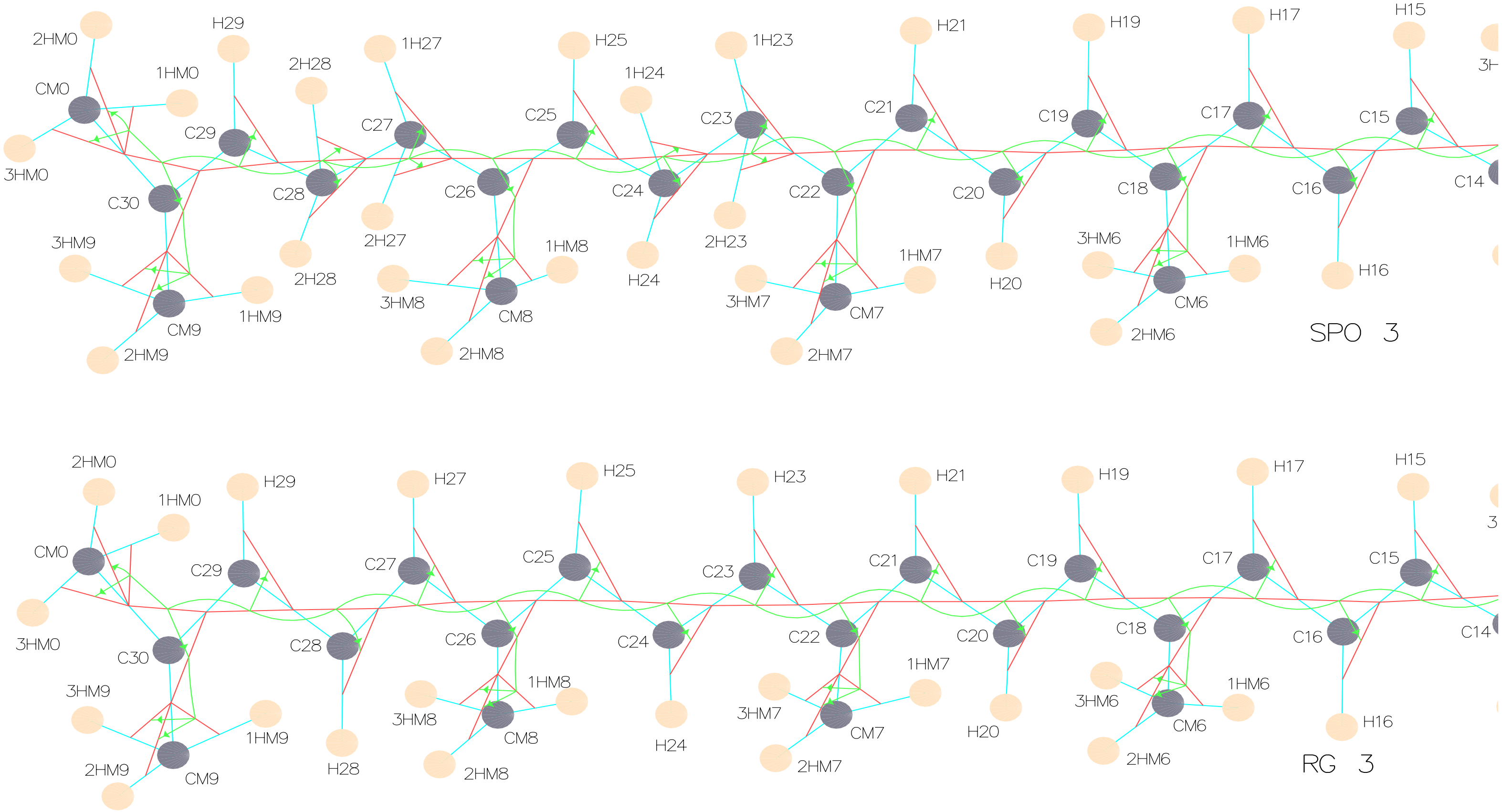


FIGURE 54. SPO 3 and RG 3

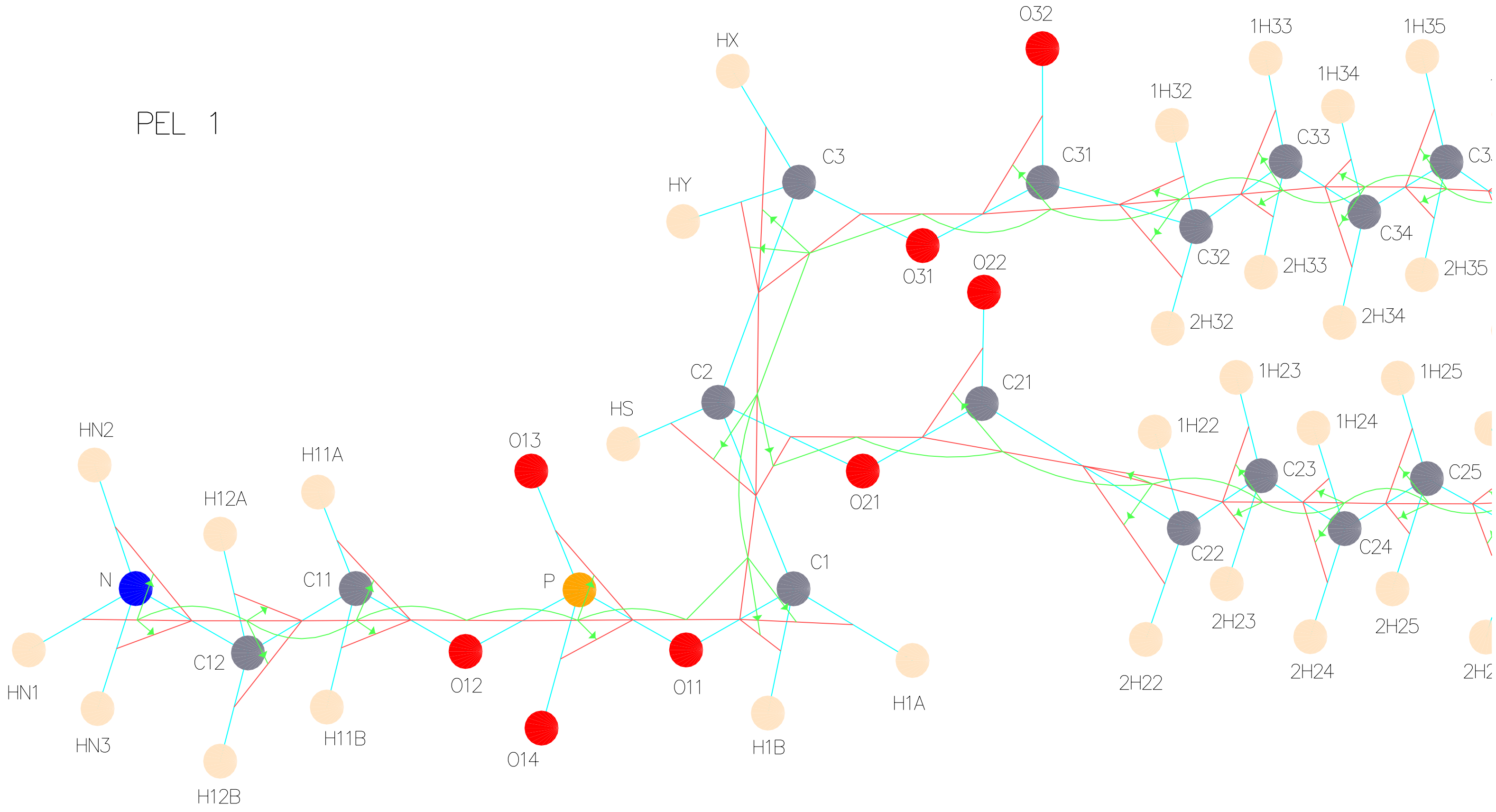


Figure 55. PEL 1

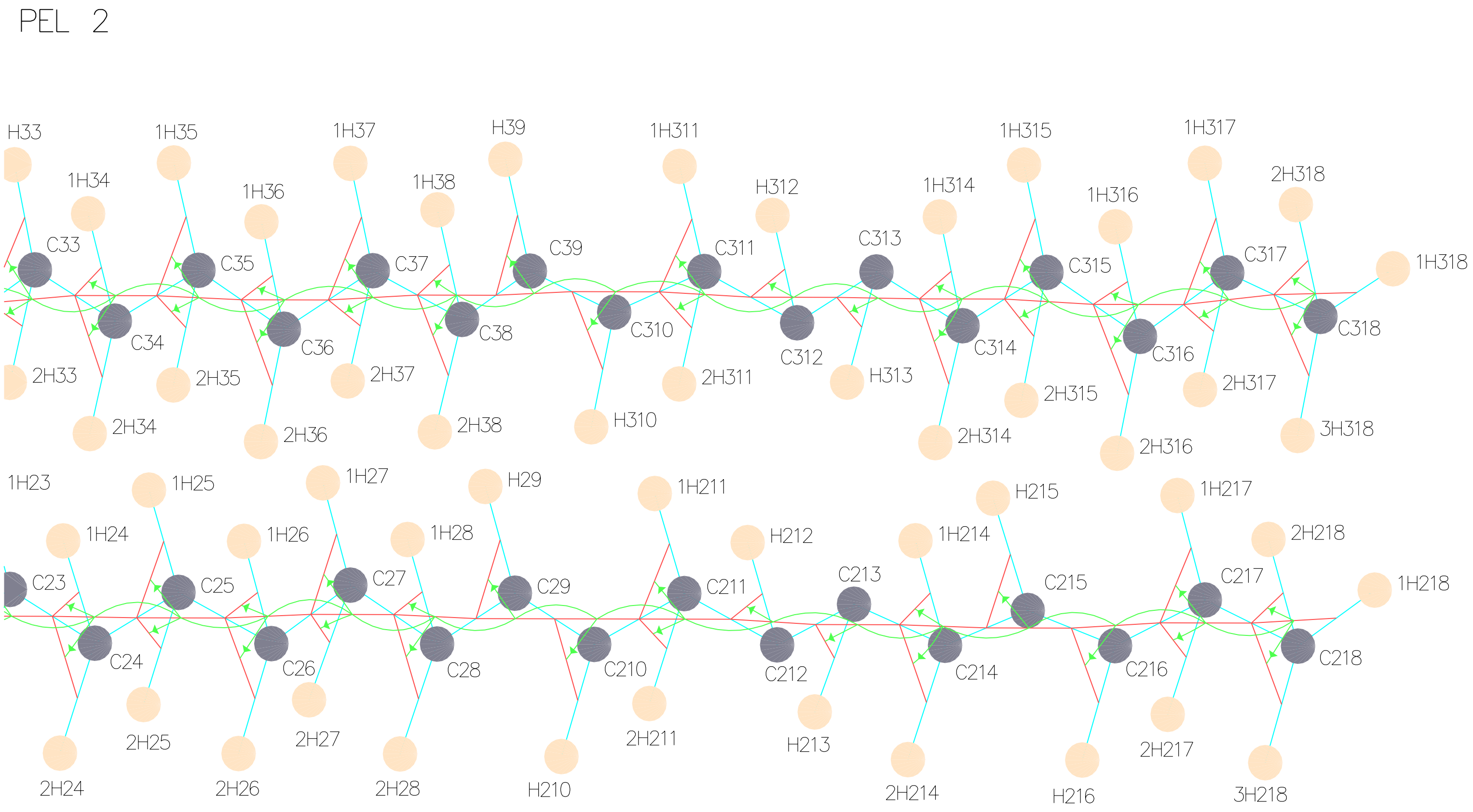


Figure 56. PEL 2

Glossary

The definitions for some biochemical terms given here are only in the light of what is discussed in this thesis; thus, they may be simplified and not give all the depth or the aspect of the word used in this or different field.

α-helix A type of helical structure of backbone of protein molecules where one turn of the helix is made by a little over four amino acids.

χ (wedge) angle A free rotation (wedge) angle around the main-chain of a side chain; χ_1 is for the rotation around the bond $\{CA, CB\}$, χ_2 is for the rotation around the bond $\{CB, CG\}$, and so on.

acidic A state of environment having more hydrogen ion, H^+ present than neutral (pH=7).

amino acid A class of molecules that make up proteins; a basic building block of proteins.

amino group A group of atoms with a nitrogen atom to which three hydrogen atoms are bonded; thus it has a positive charge as a group; -NH_3.

apoprotein A protein that is a part of a complex of proteins and other molecules.

backbone A part of an amino acid which is repeated in exactly the same manner with some modification for that of proline; or a chain of such parts; sometimes only refers to the nitrogen and carbon atoms which form the main chain of the backbone.

bacteriochlorophyll (BCL) A pigment in a photosynthetic unit (PSU) which absorbs light at a wavelength of around 800 nm to 880 nm; several different types are known according to the wavelength absorbed; also, according to the chemical structure.

basic A state of environment having less hydrogen ions, H^+, present than neutral (pH=7).

C terminus The end of a peptide or protein for which a carbon atom (C) is the terminal atom of the main-chain of the backbone.

carboxyl group A group of atoms of a carbon atom with two oxygen atoms bonded to it thus has a negative charge as a group; -COO.

carotenoid A light absorbing pigment in a photosynthetic unit (PSU) which absorbs light at a wavelength of around 500 nm; also, protects the BCLs from being in a certain dangerous state.

conjugated bonds A sequence of alternating double and single bonds.

covalent bond A bond between two atoms resulting from the sharing of electrons.

conserved Amino acids of the corresponding residues being the same for two specific species.

cytoplasm The fluid contents of the cell inside of a cell membrane.

heterodimer A pair of α- and β-apoproteins joined together to form a unit.

homologous Having similar amino acid sequences.

homology modeling Making a model of a protein from one species using the protein from the other species based on sequence homology.

hydrocarbon A molecule or a part of a molecule which consists only of carbon and hydrogen atoms.

hydrogen bond A bonding between a hydrogen atom and another element; weaker than the bonding of covalent bond; plays a major role in determining the properties of proteins and other compounds.

hydrophilic A characteristic of being ready to interact with water; strongly polar and favoring interaction with water.

hydrophobic A tendency to avoid interaction with water molecules; very stable.

in vivo Inside real, living cell.

ligate To be connected noncovalently or attracted by chemical forces.

light-harvesting complex (LHC) A system in a photosynthetic unit (PSU) which is responsible for absorbing light and transferring the energy toward the reaction center (RC).

N terminus The end of a peptide or protein for which a nitrogen atom (N) is the end atom of the backbone main-chain.

noncovalent bond A chemical attraction between two atoms without being covalently bonded.

peptide A (short) chain of amino acids.

peptide bond The bond connecting two amino acids; the rotation around this bond is usually either 0° or 180°.

periplasm The area outside of the cell membrane.

photosynthesis A process in which an organism such as a plant or a bacterium, using sunlight energy, converts carbon dioxide (CO_2) and water (H_2O) to carbohydrate, used as energy source in animals and plants, producing oxygen and water.

photosynthetic unit (PSU) A unit of systems which is responsible for carrying out photosynthesis; includes a reaction center (RC) and light-harvesting complexes (LHCs); corresponds to the photosystem in plants.

polar A characteristic of the bonding of two atoms where the electrons are not equally shared, resulting in one atom being slightly positively charged and the other slightly negatively charged.

primary structure Amino acid sequences of protein molecules.

protomer complex (PC) A subunit from which a bigger structure is built; in LHC, a unit of the heterodimer and the chromophores which is repeated to complete the ring of LHC.

protonated A state of having an extra proton, resulting in being charged positively.

reaction center (RC) A part of the photosynthetic unit where the actual reaction of photosynthesis takes place which uses the light energy captured and transferred from the light-harvesting complex.

residue The part of an amino acid which is left after forming a protein chain.

rotamer A statistically obtained, common pattern of χ angles.

secondary structure A structure of backbone of protein molecules.

sequence alignment The arrangement of sequences of amino acids in such a way as to align areas that share common properties; or to align so that the number of amino acids which agree between the sequences would be the highest.

sequence homology Refers to the situation where the sequences of amino acid types are very similar.

side chain In an amino acid, a part which is bonded to the C^α carbon atom; varies for each amino acid; also used in other molecules as a part which comes off from the main group of atoms.

transmembrane Through or across a membrane.

van der Waals' sphere A sphere occupied by the electrons in an atom.

Bibliography

1. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *The Protein Data Bank*, Nucleic Acids Research **28** (2000), 235–242.
2. B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, *The Swiss-Prot Protein Knowledgebase and Its Supplement TrEMBL in 2003*, Nucleic Acids Research **31** (2003), 356–370.
3. B. Bollobás, *Graph Theory*: *An Introductory Course*, Graduate Texts in Mathematics, Springer-Verlag New York, Inc., New York, 1979.
4. C. Branden and J Tooze, *Introduction to Protein Structure*, Barland Publishing, Inc., New York, 1991.
5. R. A. Brunisholz, I. Bissig, E. Niederer, F. Suter, and H. Zuber, *Structural Studies on the Light-Harvesting Polypeptides of Rp. acidophila*, Progress in Photosynthesis Research (J. Biggins, ed.), vol. II.1, Martinus Nijhoff Publishers, Dordrecht, The Netherland, 1987, pp. 13–16.
6. *The CancerWEB*, the Dept. of Medical Oncology, University of Newcastle upon Tyne, 2003, http://cancerweb.ncl.ac.uk/omd/index.html.
7. W. Davis, *MolProbity*, 2003, http://kinemage.biochem.duke.edu/molprobity/index-jm.html.
8. D. Dix, *Polyspherical Coordinate Systems on Orbit Spaces with Aplications to Biomolecular Comformation*, 2003, http://www.math.sc.edu/ dix/coord.pdf.
9. ______, *Geometric Formulae for Five-Membered Ring* (in preparation).
10. D. Dix and S. Johnson, *IMIMOL*, 2003, http://www.math.sc.edu/ dix/.
11. M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*, 3rd ed., Wiley, New York, 2000.
12. X. Hu, A. Damjanović, T. Ritz, and K. Schulten, *Architecture and mechanism of the light-harvesting apparatus of purple bacteria*, Proc. Natl. Acad. Sc.i U. S. A. **95** (1998), 5935–5941.

13. X. Hu, T. Ritz, A. Damjanović, F. Autenrieth, and K. Schulten, *Photosynthetic apparatus of purple bacteria*, Quarterly Reviews of Biophysics **35** (2002), 1–62.

14. X. Hu and K. Schulten, *Model for the Light-Harvesting Complex I (B875) of Rhodobacter sphaeroides*, Biophysical Journal **75** (1998), 683–694.

15. C. N. Hunter, J. D. Pennoyer, J. N. Sturgis, D. Farrelly, and R. A. Niederman, *Oligomerization states and associations of light-harvesting pigment-protein complexes of Rhodobacter sphaeroides as analyzed by lithium dodecyl sulfate-polyacrylamide gel electrophoresis*, Biochemistry **27** (1988), 3459–3467.

16. J. Koepke, X. Hu, C. Muenke, K. Schulten, and H. Miche, *The crystal structure of the light-harvesting complex II (B800-850) from Rhodospirillum molischianum*, Structure **4** (1996), 581–597.

17. A. R. Leach, *Molecular Modeling*, 2nd ed., Pearson Education Limited, New York, 2000.

18. **D. R. Lide and H. P. R. Frederikse (eds.),** *CRC Handbook of Chemistry and Physics*, 76th ed., CRC Press, Inc., New York, 1996.

19. S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson, *The Penultimate Rotamer Library*, Proteins **40** (2000), 389–408.

20. Jr. MacKerell A. D., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. N Ngo, D. T. guyen, B. Prodhom, III Reiher W. E., B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, *All-atom empirical potential for molecular modeling and dynamics: Studies of proteins*, Journal of Physical Chemistry B **102** (1998), 3586–3616.

21. J. L. Markley, A. Bax, Y. Arata, C. W. Hilbers, R. Kaptein, B. D. Sykes, P. E. Wright, and K. Wüthrich, *Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids*, J. Mol. Biol. **280** (1998), 933–952.

22. C. K. Mathews and K. E. Van Holde, *Biochemstry*, 2nd ed., The Benjamin/Cummings Publishing Company, Inc., Menlo Park, CA, 1996.

23. M. Z. Patpis, S. M. Prince, T. Howard, R. J. Cogdell, and N. W. Isaacs, *The structure and thermal motion of the B800-850 LH2 complex from Rps.acidophila at 2.0 Åresolution and 100K: new structural features and functionally relevant motions*, J Mol Biol. **326** (2003), 1523–1538.

24. S. M. Prince, M. Z. Papiz, A. A. Freer, G. McDermott, A. M. Hawthornthwaite-Lawless, R. J. Cogdell, and N. W. Isaacs, *Apoprotein Structure in the LH2 Complex from* Rhodopseudomonas

acidophila *Strain 10050: Modular Assembly and Protein Pigment Interactions*, J. Mol. Biol. **268** (1997), 412–423.

25. T. Ritz, S. Park, and K. Schulten, *Kinetics of Excitation Migration and Trapping in the Photosynthetic Unit of Purple Bacteria*, J. of Phy. Chem. B **105** (2001), 8259–8267.
26. J. J. Rotman, *The Theory of Groups*: *An Introduction*, Second, Allyn and Bacon, Inc., Boston, 1973.
27. K. Schulten, *From Simplicity to Complexity and Back: Function, Architecture, and Mechanism of Light-harvesting Systems in Photosynthetic Bacteria*, Simplicity and Complexity in Proteins and Nucleic Acids, (H. Frauenfelder, J. Deisenhofer, and P. G. Wolynes, Eds.), Dahlem University Press, 1999, pp. 227–253.
28. R. Sayle, *RasMol*, University of Massachusetts, 2003, http://www.umass.edu/microbio/rasmol.
29. *BioTech Life Science Dictionary*, BioTech Resources Web Project, Indiana Institute for Molecular and Cellular Biology, The Trustees of Indiana University, 2003, http://biotech.icmb.utexas.edu/search/dict-search.html.
30. A. J. Young and Web-Writer Inc. (monitors), *Carotenoids*, International Carotenoid Society, 2002, http://www.carotenoidsociety.org/society/fsociety.html.