Group Contributions for Estimating Standard Gibbs Energies of Formation of Biochemical Compounds in Aqueous Solution

Michael L. Mavrovouniotis
Systems Research Center and Chemical Engineering, A.V. Williams Building, University of Maryland, College Park, Maryland 20742

Accepted for publication May 29, 1990

A method is presented for the estimation of the standard Gibbs energies of formation of biochemical compounds (and hence the Gibbs energies and equilibrium constants of biochemical reactions) from the contributions of groups. The method employs a large set of groups and special corrections. The contributions were estimated via multiple linear regression, using screened and weighted literature data. For most of the data employed, the error is less than 2 kcal/mol. The method provides a useful first approximation to Gibbs energies and equilibrium constants in biochemical systems.

INTRODUCTION

Need for Equilibrium Data on Biotransformations

In recent years, the detailed study and analysis of biochemical reaction systems has allowed the systematic modelling, analysis, and improvement of bioprocesses. Basic data necessary for the modelling of biochemical reactions include reaction kinetics, which depend on the source of the enzyme (i.e., the host cell), and reaction equilibria, which are determined solely from the stoichiometry of the reaction and are independent of the enzyme catalyzing the reaction.

Equilibria are useful not only for actual enzymatic reactions but also for other stoichiometrically consistent biotransformations. Knowing the overall equilibrium constant for a whole pathway, for example, allows one to determine whether the pathway is thermodynamically favorable, whether it can lead to production of energy-rich ATP from ADP, or whether additional thermodynamic activation through ATP is necessary to make the pathway feasible. The equilibrium constants are also useful for determining ranges for intracellular concentrations of metabolites; if a metabolite concentration is too high it will reverse the direction of the reactions that produce the metabolite.

The thermodynamic analysis can be carried out using the standard Gibbs energy of reaction, $\Delta G^{0\prime}$ which is closely related to the equilibrium constant, K:

$$-\Delta G^{0\prime} = RT \ln K \tag{1}$$

where R is the ideal-gas constant and T is the temperature. The standard state for $\Delta G^{0\prime}$ is a dilute aqueous solution at $T = 25°C$, pH 7, and concentrations of compounds (other than H^+, OH^-, and H_2O) equal $1M$. For most biochemical reactions, experimental data on equilibrium constants or Gibbs energies are not available. However, the standard Gibbs energy of reaction is related to the standard Gibbs energies of formation of its reactants and products. Letting V_i be the stoichiometric coefficient of compound S_i, we can write a reaction (or any transformation with known stoichiometry) as:

$$\sum V_i S_i = 0 \tag{2}$$

where $V_i > 0$ for products and $V_i < 0$ for reactants. Let $\Delta G_i^{0\prime}$ be the Gibbs energy of formation of S_i. The Gibbs energy of reaction, $\Delta G_i^{0\prime}$, is then given by the equation:

$$\Delta G^{0\prime} = \sum V_i \Delta G_i^{0\prime} \tag{3}$$

For most biochemical compounds, Gibbs energies of formation in aqueous solution are not known *a priori*. Note, however, that from Gibbs energies of formation of a set of compounds one can calculate the Gibbs energy for *any* chemical or biochemical transformation (in aqueous solution) whose reactants and products are in this set of compounds.

Group-Contribution methods[1-7] have been widely used to estimate numerical values of thermodynamic properties of pure compounds, especially at the ideal-gas state. To estimate the property for a particular compound, one views the compound as composed of functional groups and sums amounts contributed by

Biotechnology and Bioengineering, Vol. 36, Pp. 1070–1082 (1990)

CCC 0006-3592/90/01001070-013$04.00

each group to the overall value of the property (or a suitable mathematical function of the property).

A given group-contribution method targets a specific thermodynamic property (at some standard state), for a class of compounds. The method must provide a set of functional groups, which serve as the building blocks for the compounds of interest. The contribution of each group to the thermodynamic property of interest must also be provided, along with the *origin,* a starting value that is used in the estimation (and is constant for all compounds). To estimate the property of a particular compound, one decomposes the compound into groups, and adds the contributions of the groups to the constant origin. If a certain group occurs several times within the compound, then its contribution must be multiplied by its number of occurrences. For some properties, additional algebraic operations (including corrections for special structural features of the compound) must be performed on the summation of the contributions to obtain the final value.[4]

Let C_o be the origin for the property C, and let C_i be the contribution of group g_i which is used N_i times in the compound. The property C for the whole compound is calculated as:

$$C = C_o + \sum N_i C_i \qquad (4)$$

In some cases, additional contributions must be included to account for other features of the molecule as a whole, such as aromaticity, special interactions among groups, etc.

A group contribution method can be developed using data (i.e., the value of the property for several compounds) to estimate the contributions of the groups to the property of interest. In effect, if the data consist of values of C for a set of compounds, and the molecular structures (hence, the N_i values) of the compounds are known, then a number of equations of the form of eq. (4) are available, and the unknown origin, C_o, and contributions, C_i, can be determined. In principle, one only needs as many independent data points as there are unknowns. However, errors in the data might lead to very large errors in the contributions. Also, eq. (4) is only an approximation and the contributions would be strongly biased by peculiarities of the data points used. Thus, it is generally desirable to have as many data points as possible and obtain values for the contributions by minimizing the sum of the square of the errors; the error is defined for each data point as the difference between the given value and the value estimated by eq. (4). Since eq. (4) is linear, the minimization of the errors is easily accomplished through multiple linear regression.

If C is a property applicable to reactions in the linear manner suggested by eq. (3), then data on reactions and data on compounds can be treated uniformly during both the development and the utilization of the group-contribution method. In other words, reactions can be viewed as collections of groups by subtracting the number of occurrences of each group in reactants from its occurrences in products, because both eqs. (3) and (4) are linear. Of course, groups may participate in a reaction with either positive or negative net coefficients. The same linear-combination treatment must also be applied for the contribution of the *origin* and all additional corrections.

A limited method has been recently developed for the estimation of the Gibbs energies of formation of biochemical compounds (and hence the Gibbs energies and equilibrium constants of biochemical reactions) from the contributions of groups.[5] The new method proposed here has wider applicability because it provides the contributions of a considerably larger set of groups. Through this set of groups and a few important corrections for group interactions especially critical in biochemical compounds, the proposed method also achieves better accuracy. Although the method was motivated by the need for thermodynamic analysis of biochemical compounds and reactions, it can be used to estimate the standard Gibbs energy and the equilibrium constant of any organic reaction taking place in an aqueous solution.

GROUP CONTRIBUTION METHOD

The data used in the regression were taken from several sources.[8-16] A screening of the data was carried out to eliminate particularly unreliable data. A weight reflecting a subjective judgement of accuracy was assigned to each data point, in an effort to minimize the adverse effects of less accurate data points.

A large set of groups was used, in order to cover a wider spectrum of biochemical compounds and achieve better accuracy. In addition, corrections were introduced to account for certain group interactions such as the interaction between a nitrogen and an adjacent carbonyl group (because the nitrogen has amide rather than amine properties). More global characteristics of molecules (e.g., the number of aromatic rings) were also captured by special corrections. However, entropy corrections for intramolecular symmetry were not introduced, following the approach of Joback and Reid,[4] because the complexity of biochemical compounds makes the identification of the correct symmetry numbers difficult.

Several small compounds were *not* broken down into groups. Some of these were simply not decomposable (e.g., hydrogen cation, methane, formaldehyde, many inorganic ions) while others could be decomposed but would cause large errors (e.g., oxalate, methanol). Each of these small compounds was represented as consisting of a special single group.

Special groups were also used for certain sets of complex biochemical compounds with important metabolic

roles. For example, the pair $NAD^+/NADH$ was represented by a single group, which stands for the transformation of NAD^+ to NADH, and thus represents the structural differences between the two compounds, rather than a structural feature of any one compound. This was necessary partly because of the complex structure of NAD^+ and NADH and partly because it is very important to have an accurate contribution for their interconversion, which occurs in a large number of biochemical reactions.

Throughout the development and application of the group-contribution method, all compounds and groups were represented in their common state in aqueous solution. Thus, amine groups normally have a proton attached (e.g. $R—NH_3^{1+}$ rather than $R—NH_2$), carboxylic acids are in their anion form ($R—CO—O^{1-}$ rather than R—CO—OH), amino-acids in their zwitter-ion forms, etc.

The determined contributions of groups are shown in Tables I–V. Table I shows the contributions of all open-chain groups that have only one free (single) bond; Table II shows the contributions of open-chain groups that have more than one free bond; Table III lists the contributions of groups that participate in one ring; Table IV details the contributions of groups that participate in two (fused) rings; Table V lists the contributions of the *origin* and special corrections. In addition, Table VI shows the Gibbs energies of formation of molecules that were treated as special single groups, while Table VII shows the contributions for special pairs of compounds, whose interconversion is important in the metabolism.

Because for many biochemical compounds that contain rings the decomposition into groups and the application of the appropriate corrections are especially cumbersome, the Gibbs energies of formation of a large number of cyclic compounds were precomputed (using the developed method) and are listed in Table VIII.

Table I. Contributions of groups with one free single bond.

Group	Contribution (kcal/mol)
$—CH_3$	8.5
$—NH_2$	11.0
—OH (attached to benzene aromatic ring)	−31.2
—OH (primary)	−28.6
—OH (secondary)	−31.4
—OH (tertiary)	−29.9
$—NH_3^{1+}$	4.8
—CH=O	−17.3
—SH	14.0
$—COO^{1-}$	−71.4
$—PO_3^{2-}$	10.2
$—SO_3^{1-}$	−105.1
$—CO—OPO_3^{2-}$	−72.1
$—OPO_3^{2-}$ (primary)	−28.8
$—OPO_3^{2-}$ (secondary)	−29.4
$—OPO_3^{2-}$ (tertiary)	−24.9

Table II. Contributions of groups with two or more free bonds, not participating in a ring.

Open-chain group	Contribution (kcal/mol)
≡C—	24.0
≡N	15.5
≡CH	36.3
=C<	4.4
—CH=	11.0
$=CH_2$	19.0
$=NH_2^{1+}$	0.8
=NH	14.3
>C<	−14.0
—CH<	−5.4
>N—	5.8
$—NH^{1+}<$	8.0
$—CH_2—$	1.7
>CO	−27.3
$—O—PO_2^{1-}—$	−5.2
—S—S—	5.9
$—CO—O^{1-}$	−73.6
—S—	9.5
$>NH_2^{1+}$	6.9
—O— (attached to $—O—PO_2^{1-}—$ group)	−24.6
—O—	−22.5
—NH—	7.8

EXAMPLES

A few example calculations will be provided here to illustrate the use of the group-contribution method. Consider the estimation of the Gibbs energy of formation of glutamate, whose syntactic formula is shown in Figure 1(a). It can be broken down into groups in a straightforward manner, as shown in Figure 1(b). As shown in Table IX, the calculation entails the addition of the contributions (multiplied by the number of occurrences of each group) to the fixed contribution of the origin; the latter is taken from Table V. In this example, no special corrections are needed. The final result is −164.7 kcal/mol, which deviates by 2.4 kcal/mol from the literature value[8] of −167.1 kcal/mol.

As an example involving a complex cyclic compound, consider next the estimation of the Gibbs energy of formation of ATP, whose structure is shown in Figure 2. The structure is broken down into groups in Figure 3. Note that it is necessary to classify some substituents—the hydroxyl and phosphate goups—as primary, secondary, or tertiary. Also, the participation of bonds in rings must be taken into consideration. Apart from the groups the correction for two heteroaromatic rings, which are identified in Figure 3, must be applied; as noted in Table V, a heteroaromatic ring has, by resonance, $4n + 2$ delocalized π-electrons, in accordance with Hückel's rule.[17] Table X shows the calculation of the Gibbs energy from the contributions. The result is −60.4 kcal/mol, equal to the precomputed value −60.4 kcal/mol given in Table VIII. However, the values in Table VIII were calculated *before* rounding off the contributions to one decimal place, and if one com-

Table III. Contributions of groups participating in one ring. The rings in which the compounds participate are *not* aromatic, unless it is otherwise indicated.

Ring group	Contribution (kcal/mol)
>N=$^{1+}$ (a single and a double bond in a non-aromatic ring)	−0.1
>C— (in one benzene ring)	1.1
>C= (two single bonds in a nonbenzene ring)	22.8
>C= (a single bond and a double bond in a non-benzene ring)	7.9
>C<	−13.7
=N—	10.5
>CH (in one benzene ring)	8.6
—CH=	9.6
—N<	7.4
—CH<	−2.6
—O—CO—	−54.3
—CO—	−27.4
—O—	−24.1
—NH—	9.9
—CH_2—	6.3
—O—PO_2^{1-}—	15.6

Table IV. Contributions of groups participating in two fused rings.

Two-ring group	Contribution (kcal/mol)
>C— (in two fused benzene rings)	2.3
—N< (in two fused nonbenzene rings)	19.1
—CH< (in two fused nonbenzene rings)	−1.3
>C= (in two fused nonbenzene rings)	16.9

putes the Gibbs energies using the rounded contributions of Tables I–V, the result may differ from the one provided in Table VIII (in the first decimal position). It was deemed appropriate to maintain only one decimal place in the contributions, to avoid any misconceptions about the accuracy of the method. Essentially, even the first decimal place of a result is not meaningful; but if it were dropped from the contributions compounds with a large number of groups could accumulate unacceptably large roundoff errors.

The identification of groups, based on the participation of their bonds in rings, and the incorporation of the appropriate corrections is clearly not a trivial matter in the case of ATP and other cyclic compounds. For this reason, the Gibbs energies of common cyclic compounds were computed and are listed in Tabel VIII. The energies of Table VIII serve as useful *reference points,* when the Gibbs energy of a compound similar to one given in the table is desired. One needs to identify only the *differences* between the compound at hand and the similar compound listed in the table, and add and subtract appropriate contributions (and corrections) that correspond to these differences. As an example, consider the calculation of the Gibbs energy of formation of ADP, with the Gibbs energy of ATP known. The comparison of the structures of the two compounds in Figure 4 indicates that ATP simply has an additional —OPO_2^{1-}— group. Thus, the Gibbs energy of ADP can be obtained from ATP by subtracting the contribution of this group. The result is obtained as $-60.4 - (-5.2) = -55.2$ kcal/mol. For a more complex example of this method of comparing compounds, the Gibbs energy of formation of 7,8-dihydro-ATP (Fig. 5) is derived in Table XI. Figure 5 shows the two structures and the groups in which they differ; two groups are replaced by two different ones, and a heteroaromatic ring correction must be removed. The calculation in Table XI shows that the result, −67.2 kcal/mol, is rather easy to obtain.

Table V. The origin and additional corrections to the Gibbs energy of formation in aqueous solution. A ring is considered heteroaromatic if it has, by resonance, delocalized π-electrons that follow Hückel's $4n + 2$ rule.[17] Note that this determination will not have to be made frequently because we have precomputed the Gibbs energy for many ring compounds (Table VIII).

Correction	Contribution (kcal/mol)
Origin	−24.7
Hydrocarbon molecule (i.e., containing only C and H)	4.0
Each aromatic ring	−6.0
Each three-carbon ring	29.5
Each amide (i.e., each nitrogen attached to a carbonyl group)	−10.5
Each heteroaromatic ring (containing nitrogen, sulfur, or oxygen)	−5.9

Table VI. Gibbs energies of small compounds that were not broken down into groups.

Formula	Name	Gibbs energy (kcal/mol)
$^{1-}OCO-COO^{1-}$	oxalate	−159.8
CH_4	methane	−8.2
$HCOO^{1-}$	formate	−84.9
$CH_2=O$	formaldehyde	−36.1
H_2CO_3	hydrated carbon dioxide	−148.9
HCO_3^{1-}	bicarbonate	−140.2
$HP_2O_7^{3-}$	pyrophosphate	−57.4
HPO_4^{2-}	phosphate	−59.6
SO_3^{2-}	sulfite	−118.4
NO_3^{1-}	nitrate	−27.5
NO_2^{1-}	nitrite	−7.0
NH_4^{1+}	ammonium	−18.1
H_2O	water	−56.6
OH^{1-}	hydroxide anion	−47.1
H^{1+}	hydrogen cation	−9.5

Table VII. Gibbs energy differences for special pairs of compounds.

Group	Contribution (kcal/mol)
NADH (reduced) *minus* NAD^+ (oxidized)	4.74
NADPH (reduced) *minus* $NADP^+$ (oxidized)	4.74
Substituted Coenzyme-A *minus* SH-Coenzyme-A (to calculate the Gibbs energy difference, the contributions of the substituent's groups must be added to the correction)	−13.2
Pyocyamine reduced *minus* Pyocyamine oxidized	−17.8

Table VIII. Estimated standard Gibbs energies of formation of selected classes of cyclic compounds. Within each class, the compounds are in increasing order of total number of atoms.

Compound	Standard Gibbs energy of formation in aqueous solution (kcal/mol)
Derivatives of coenzyme A (relative Gibbs energies)[a]	
Acetyl-coenzyme-A	−33.1
Oxalyl-coenzyme-A	−113.0
Acrylyl-coenzyme-A	−11.6
Propionyl-coenzyme-A	−31.4
Malonyl-coenzyme-A	−111.3
Crotonoyl-coenzyme-A	−11.0
Acetoacetyl-coenzyme-A	−58.6
Keto-butyryl-coenzyme-A	−58.6
Butyryl-coenzyme-A	−29.7
Methylmalonyl-coenzyme-A	109.8
Succinyl-coenzyme-A	−109.6
Alanyl-coenzyme-A	−33.4
3-Hydroxy-butyryl-coenzyme-A	−68.1
Keto-hexanoyl-coenzyme-A	−55.2
Hexanoyl-coenzyme-A	−26.3
3-Hydroxy-hexanoyl-coenzyme-A	−64.7
Hydroxy-crotonyl-coenzyme-A	−71.6
Palmitoyl-coenzyme-A	−9.2
Nucleosides and bases	
Uracil	−67.4
4,5-Dihydro-uracil	−68.1
Hypoxanthine	19.4
Inosine	19.4
Thymine	−60.6
Adenine	76.8
Xanthine	−28.7

Table VIII. *(continued)*

Guanine	28.7
Uric acid	−87.3
Orotate	−140.5
Cytosine	−10.0
Thymidine	−146.9
2′-Deoxyadenosine	−9.5
Adenosine	−49.7
Guanosine	−97.8
Nucleotides	
2′-Deoxycytidine monophosphate (d-CMP)	−96.4
Uridine monophosphate (UMP)	−194.1
3′,5′-Cyclic adenosine monophosphate (cAMP)	9.0
Cytidine monophosphate (CMP)	−136.7
3′,5′-Cyclic guanosine monophosphate (cGMP)	−39.1
2′-Deoxythymidine monophosphate (d-TMP)	−147.0
Inosine monophosphate (IMP)	−107.3
Thymidine monophosphate (TMP)	−187.3
Adenosine monophosphate (AMP)	−49.9
Guanosine monophosphate (GMP)	−97.9
2′-Deoxycytidine diphosphate (d-CDP)	−101.7
Uridine diphosphate (UDP)	−199.4
Cytidine diphosphate (CDP)	−141.9
2′-Deoxythymidine diphosphate (d-TDP)	−152.3
Inosine diphosphate (IDP)	−112.6
Thymidine diphosphate (TDP)	−192.5
Adenosine diphosphate (ADP)	−55.1
Guanosine diphosphate (GDP)	−103.2
Uridine triphosphate (UTP)	−204.6
Cytidine triphosphate (CTP)	−147.2
Inosine triphosphate (ITP)	−117.8
Thymidine triphosphate (TTP)	−197.8
Adenosine triphosphate (ATP)	−60.4
Guanosine triphosphate (GTP)	−108.4
Derivatives of nucleosides and nucleotides	
UDP-glucuronate	−402.9
UDP-glucose	−358.4
CDP-glucose	−301.0
GDP-galactose	−262.3
GDP-glucose	−262.3
GDP-mannose	−262.3
Adenosyl homocystein	−80.1
Adenylosuccinate	−199.6
Adenylyl-α-aminoadipate	−166.8
Tetrahydrofolate and derivatives	
7,8-Dihydrofolate	−74.1
Tetrahydrofolate	−85.2
N^5,N^{10}-Methenyl-tetrahydrofolate	−68.5
N^5,N^{10}-Methylene-tetrahydrofolate	−64.2
N^{10}-Formyl-tetrahydrofolate	−114.8
N^5-Formyl-tetrahydrofolate	−115.4
N^5-Methyl-tetrahydrofolate	−79.1
N^5-Formimino-tetrahydrofolate	−75.9
Cyclic sugars	
Arabinose	−178.3
Lyxose	−178.3
Ribose	−180.1
Ribulose	−178.5
Xylose	−178.3
Xylulose	−178.5
6-Deoxygalactose	−178.6
Fuculose	−181.3
Rhamnose	−178.6
Rhamnulose	−181.3
Allose	−214.1
Altrose	−214.1
Fructose	−216.7
Galactose	−214.1

Table VIII. *(continued)*

Glucose	−214.1
Gulose	−214.1
Idose	−214.1
Inositol	−228.4
Mannose	−214.1
Psicose	−216.7
Sorbose	−216.7
Tagatose	−216.7
Talose	−214.1
Cellobiose	−363.2
α-lactose	−363.2
β-Maltose	−363.2
Sucrose	−367.2
Trehalose	−363.2
Cellotriose	−512.2
Sugar phosphates	
2'-Deoxyribose-5'-phosphate	−140.1
Arabinose-5-phosphate	−180.3
Ribose-1-phosphate	−178.2
Ribose-5-phosphate	−180.3
Fuculose-1-phosphate	−182.0
Rhamnulose-1-phosphate	−182.0
Glucuronate-1-phosphate	−256.6
Fructose-1-phosphate	−216.9
Fructose-6-phosphate	−216.9
Galactose-1-phosphate	−212.1
Glucose-1-phosphate	−212.1
Glucose-6-phosphate	−214.3
Mannose-1-phosphate	−212.1
Mannose-6-phosphate	−214.3
Fructose-1,6-diphosphate	−217.1
Glucose-1,6-diphosphate	−212.3
Sucrose-6-phosphate	−367.4
Hydrocarbons	
Cyclopropane	27.6
Benzene	31.1
Cyclo-hexane	16.9
Naphthalene	52.9
Simple alcohols, aldehydes, and ethers	
Furane	−16.4
Tetrahydrofurane	−23.7
Phenol	−11.6
1,4-Dioxane	−47.8
Benzaldehyde	2.3
Methyl phenyl ketone	0.9
Cyclohexanol	−27.3
Other compounds	
Pyridine	27.8
Niacine	−45.2
δ-Pyrroline-5-carboxylate	−66.0
Nicotinamide	−0.6
Pyrrolidone carboxylate	−114.1
Urocanate	−32.5
Creatinine	−6.9
Dihydroorotate	−148.3
2,3-Dihydro-dipicolinate	−126.1
5-Dehydroshikimate	−167.6
1-Piperidine-2,6-dicarboxylate	−132.8
Histidine	−53.4
Shikimate	−174.1
Gluconic acid γ-lactone	−213.2
Gluconic acid δ-lactone	−210.4
Histidinol	−9.0
5-Dehydroquinate	−222.5
Fructuronate	−261.2
Glucuronate	−258.5
Imidazole acetol phosphate	−35.8

Table VIII. (*continued*)

Phenylalanine	−50.7
Quinate	−229.0
Tyrosine	−89.4
δ-Gluconolactone-6-phosphate	−210.5
Histidinol phosphate	−9.1
7,8-Dihydrobiopterin	−25.6
5,6,7,8-Tetrahydrobiopterin	−36.7
Pteroate	27.2
Pteroyl glutamate	−76.1

[a] The energies listed for Co-A derivatives use coenzyme A as a reference point. In effect, the energy shown is equal to the Gibbs energy of each derivative minus the Gibbs energy of coenzyme A.

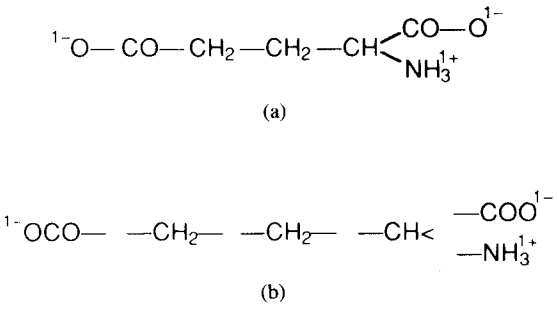


Figure 1. The structure of glutamate (a) and decomposition of the structure into groups (b).

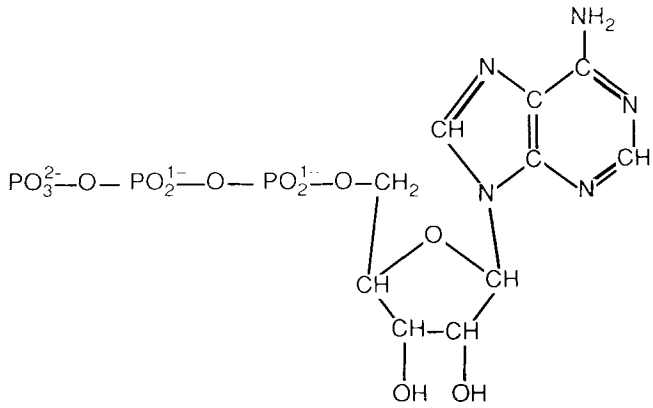


Figure 2. The structure of ATP.

A last example is provided in Figure 6 and Table XII for a biochemical reaction, catalyzed by *alcohol dehydrogenase*. The reaction is decomposed into groups in Figure 6; note that the pair NADH/NAD^+ is considered a single group. The calculation in Table XII ignores the contributions of the *origin* and the group $-CH_3$, because they are the same for ethanol and acetaldehyde, i.e., they have net number of occurrences equal to zero. The result is 4.8 kcal/mol, which compares well with literature values[9,12] of 5.5 and 5.9 kcal/mol.

DISCUSSION

Using the contributions and corrections of Tables I–VII, one can estimate the standard Gibbs energy of formation of a biochemical compound (or any organic compound) in aqueous solution, provided that the molecular structure of the compound (specifically in its form most prevalent in water) is known. The standard Gibbs energy and the equilibrium constant of a biochemical reaction (or other organic reaction) can be estimated from its stoichiometry and the molecular structures of its reactants and products.

The method has broad applicability because it provides the contributions for a comprehensive set of groups. Even though many of the compounds of Table VIII contain complex ring systems with heteroatoms, the presented group contribution technique was able to estimate their Gibbs energies. For the datapoints used in the regression, the distribution of errors is given in Table XIII and Figure 2. The error for 85% of the data points is less than 2 kcal/mol; for 95% of

Table IX. Calculation of the Gibbs energy of formation of glutamate from contributions of groups.

Group or correction	Number of occurrences	Contribution (kcal/mol)	Source	Total contribution
Origin	1	−24.7	Table V	−24.7
$-NH_3^{1+}$	1	4.8	Table I	4.8
$-COO^{1-}$	2	−71.4	Table I	−142.8
$-CH_2-$	2	1.7	Table II	3.4
$-CH<$	1	−5.4	Table II	−5.4
Total				−164.7

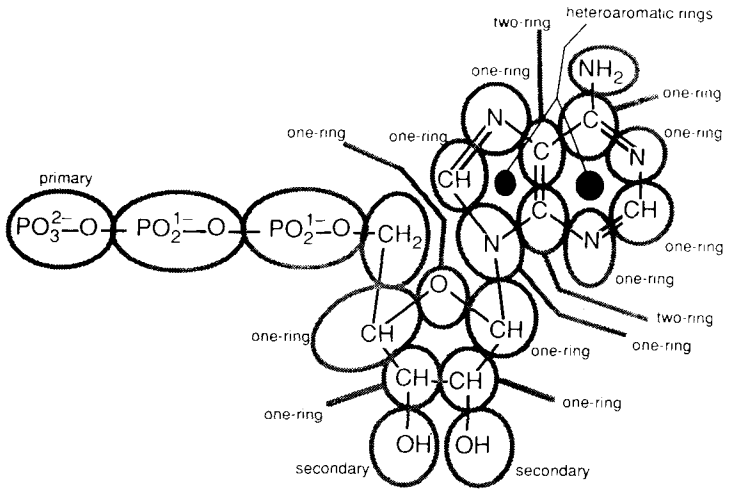


Figure 3. Decomposition of the structure of ATP into groups.

Table X. Calculation of the Gibbs energy of formation of ATP from contributions of groups.

Group or correction	Number of occurrences	Contribution (kcal/mol)	Source	Total contribution
Origin	1	−24.7	Table V	−24.7
—NH_2	1	11.0	Table I	11.0
—OPO_3^{1-}	1	−28.8	Table I	−28.8
—OH secondary	2	−31.4	Table I	−62.8
—CH_2—	1	1.7	Table II	1.7
—OPO_2^{1-}—	2	−5.2	Table II	−10.4
Ring —O—	1	−24.1	Table III	−24.1
Ring —CH<	4	−2.6	Table III	−10.4
Ring —N<	1	7.4	Table III	7.4
Ring —CH=	2	9.6	Table III	19.2
Ring =N—	3	10.5	Table III	31.5
Ring >C=	1	7.9	Table III	7.9
Two-ring >C=	2	16.9	Table IV	33.8
Heteroaromatic ring	2	−5.9	Table V	−11.8
Total				−60.4

the data, the error is less than 5 kcal/mol. Thus, the method provides an acceptable first approximation to Gibbs energies and equilibrium constants in biochemical systems.

In order to assess the sensitivity and accuracy of the group-contribution method, the following additional test was performed. In a pass through the database, each datum was eliminated with probability of 10%, using a random number generator; in effect a random portion, comprising approximately 10% of the database was eliminated. Then the contributions of the groups were estimated using only the remaining data, and, using the new contributions, the Gibbs energies of the eliminated compounds were reevaluated; the new values were compared to the original values that had been obtained using the whole database. This scheme represents one way of testing the *extrapolating* ability of the group-contribution method. The results of three such runs are shown in Table XIV and Figure 8, as the distribution of the errors (the differences between the new and the old values for the 10% portion of the database). The Gibbs energies appear quite robust in this test. For most compounds the change is less than 0.5 kcal/mol, an amount which is smaller than the accuracy of the group-contribution method. In a few cases, however, there is a change of 3 or 4 kcal/mol; this occurs when the particular set of compounds that were removed makes the estimation of the contribution of some group very difficult. For example, if there are only four datapoints on a group and two happen to be removed in the ran-

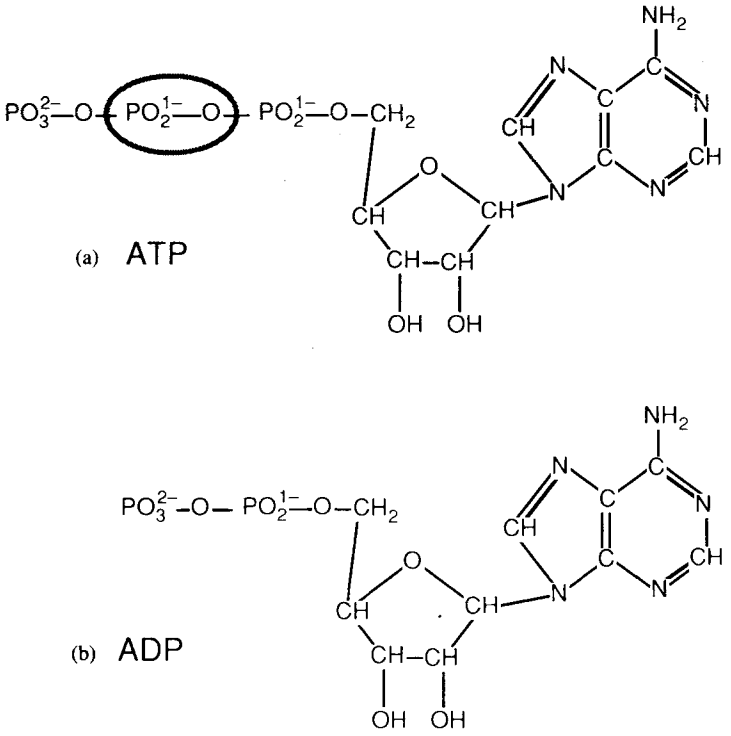


Figure 4. Comparison of the structures of ATP and ADP. The comparison indicates that ATP simply has an extra $—OPO_2^{1-}—$ group.

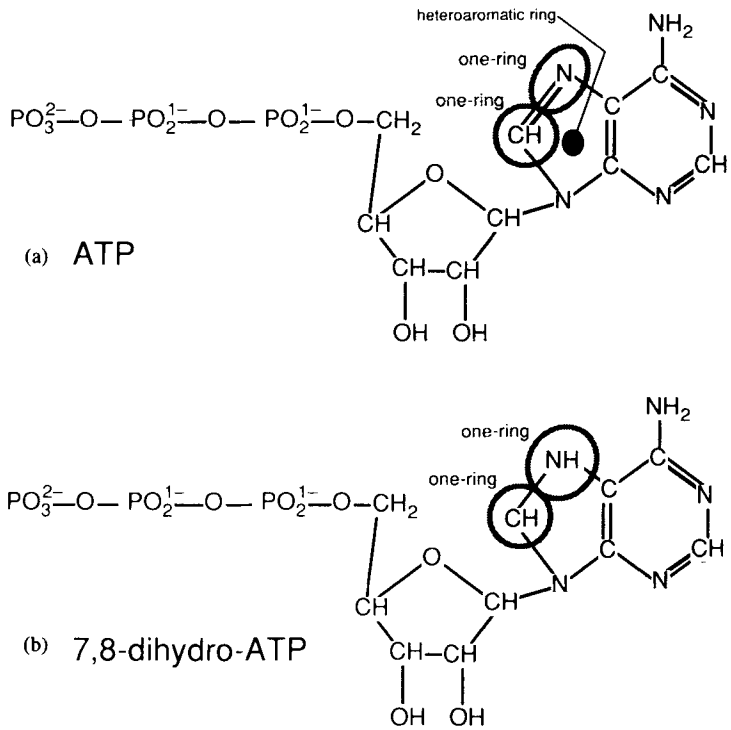


Figure 5. Comparison of the structures of ATP and 7,8-dihydro-ATP. The comparison indicates that two groups are replaced by two others, and a heteroaromatic ring correction is lost in 7,8-dihydro-ATP.

Table XI. Calculation of the Gibbs energy of formation of 7,8-dihydro-ATP from the Gibbs energy of ATP and contributions of groups in which 7,8-dihydro-ATP differs from ATP.

Group or correction	Number of occurrences	Contribution (kcal/mol)	Source	Total contribution
ATP	1	−60.4	Table X	−60.4
Ring —CH=	−1	9.6	Table III	−9.6
Ring =N—	−1	10.5	Table III	−10.5
Heteroaromatic ring	−1	−5.9	Table V	5.9
Ring —NH—	1	9.9	Table III	9.9
Ring —CH<	1	−2.6	Table III	−2.6
Total				−67.2

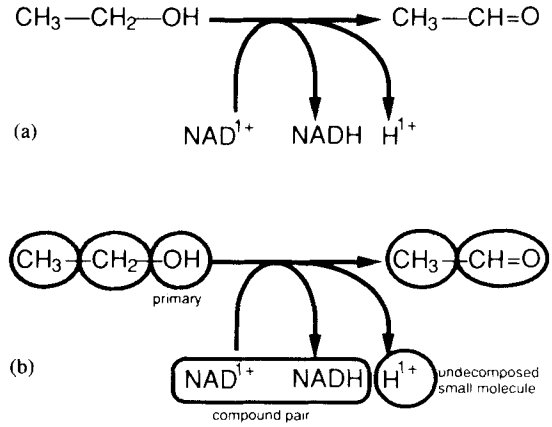


Figure 6. The reaction catalyzed by *alcohol dehydrogenase* (a) and the reaction decomposed into groups, so that its Gibbs energy can be estimated (b).

dom sweep, there is significant loss of accuracy in the contribution of the group and changes of the order of 4 kcal/mol should be expected. Note that the test that was performed entails removal of several data points simultaneously; removal of one datapoint at a time would have a much smaller effect. The most important point to note is that the error distribution of Figure 8 is much narrower than the distribution in Figure 7, suggesting that the method's performance does not deteriorate for compounds that are not present in its database.

A fundamental difficulty in the development of group-contribution methods for properties of biochemical compounds is that there are often strong interactions among groups due to *conjugation*. The *conjugates* of a compound are alternative formal arrangements of the valence electrons; a compound that is strongly influenced by conjugation cannot be adequately represented by any single structural formula and therefore cannot be properly decomposed into groups. For example, there is conjugation in the adenine portion of AMP, as shown in Figure 9. This conjugation is taken into account indirectly with a correction applicable to heteroaromatic rings (Table V). On the other hand, Figure 10 shows that in the sugar portion of AMP there are other conjugates which are not accounted for in group contributions; the problem is that there are too many combinations of group interactions and they cannot be captured with just a few special corrections.

We are currently investigating a new property-estimation framework that addresses this and other deficiencies of group-contribution methods. This framework (named ABC[18,19]) is based on using contributions of Atoms and Bonds for properties of Conjugates of a compound, and deriving the properties of the compound from properties of its conjugates. This approach has been enhanced by approximate quantum-chemical analysis and has been demonstrated for simple com-

Table XII. Calculation of the Gibbs energy of the reaction catalyzed by *alcohol dehydrogenase*, from contributions of groups.

Group or correction	Number of occurrences	Contribution (kcal/mol)	Source	Total contribution
Origin	0			
H^{1+}	1	−9.5	Table VI	−9.5
NADH *minus* NAD^+	1	4.74	Table VII	4.74
$-CH_3$	0			
$-CH_2-$	−1	1.7	Table II	−1.7
—OH primary	−1	−28.6	Table I	28.6
—CH=O	1	−17.3	Table II	−17.3
Total				−4.8

Table XIII. Distribution of the errors in the Gibbs energies of the compounds used in the estimation. The error is defined as the difference between the predicted value and the literature datum.

Error range (kcal/mol)	Number of data points
−10.0 to −9.0	1
−9.0 to −8.0	0
−8.0 to −7.0	0
−7.0 to −6.0	3
−6.0 to −5.0	4
−5.0 to −4.0	7
−4.0 to −3.0	13
−3.0 to −2.0	29
−2.0 to −1.0	53
−1.0 to 0.0	102
0.0 to 1.0	109
1.0 to 2.0	39
2.0 to 3.0	10
3.0 to 4.0	17
4.0 to 5.0	6
5.0 to 6.0	4
6.0 to 7.0	3
7.0 to 8.0	3
8.0 to 9.0	1
9.0 to 10.0	0

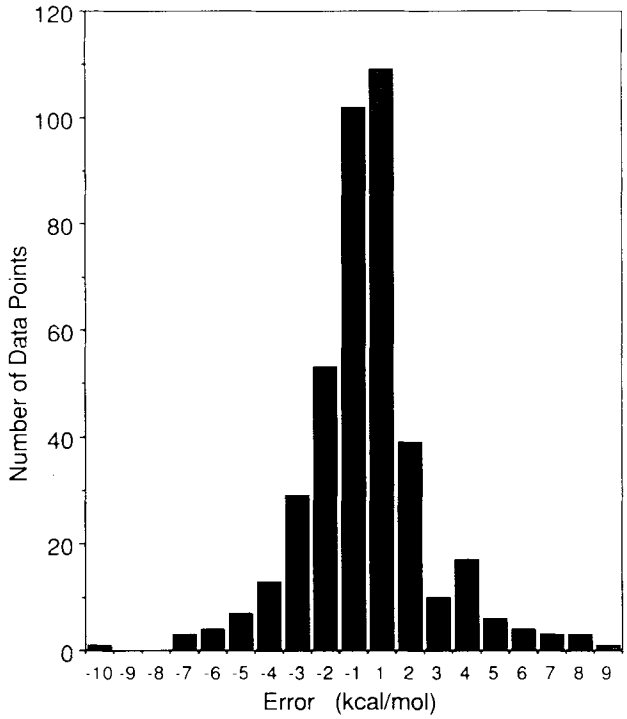


Figure 7. Distribution of errors (residuals) from the regression. Each positive label x on the horizontal axis represents errors between x − 1 and x. Each negative label −x represents errors between −x and −x + 1.

Table XIV. Distribution of the discrepancies in the Gibbs energies of compounds which were temporarily eliminated from the database to assess the sensitivity and accuracy of the group-contribution method. For each of three runs, a random (approximately 10%) portion of the database was eliminated, and a fit was performed, estimating contributions and the Gibbs energies of the omitted compounds. The distribution of the errors (i.e., the changes from the values that had been estimated using the whole database) is shown here.

Error range (kcal/mol)	Number of data points		
	Run 1	Run 2	Run 3
−3 to −2.5	1	0	0
−2.5 to −2	0	2	1
−2 to −1.5	1	0	0
−1.5 to −1	0	1	0
−1 to −0.5	3	4	0
−0.5 to 0.5	11	6	11
0.5 to 1	7	8	16
1 to 1.5	1	2	1
1.5 to 2	0	1	4
2 to 2.5	2	0	0
2.5 to 3	0	0	0
3 to 3.5	1	0	0
3.5 to 4	0	1	1
4 to 4.5	0	0	0
4.5 to 5	2	0	0

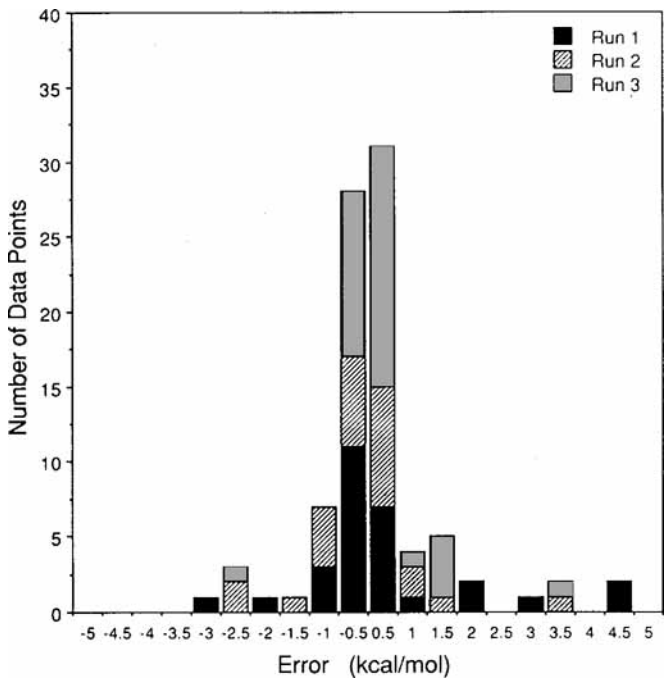


Figure 8. Distribution of errors from three runs in which a small random portion of the database was eliminated, and a fit was performed, yielding new Gibbs energy values for the omitted compounds. The distribution of the changes in the Gibbs energies is shown here (see also Table XIV). Each positive label x on the horizontal axis represents errors between x − 1 and x. Each negative label −x represents errors between −x and −x + 1.

pounds in the ideal-gas state.[18,19] Since biochemical compounds are strongly influenced by conjugation, it is expected that the ABC framework will be of great value in estimating their properties.

This research was supported in part by the Minta Martin Research Fund, the General Research Board, and the Engi-

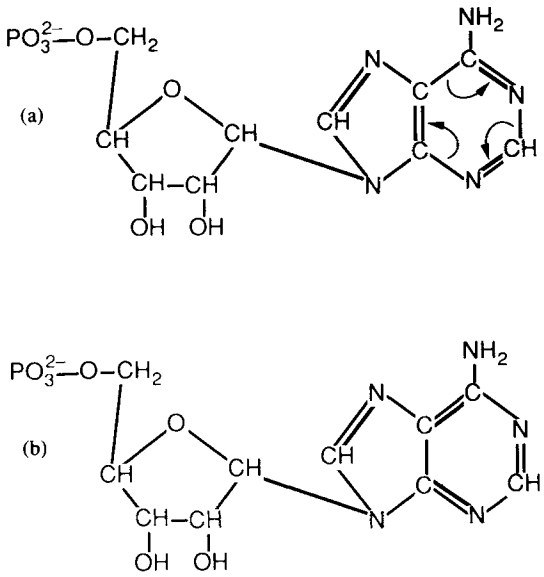
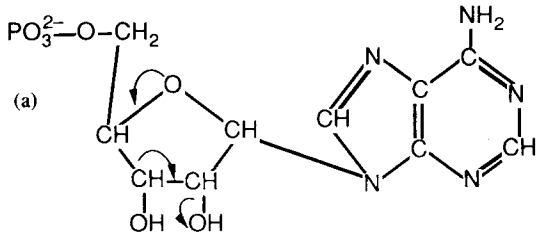


Figure 9. Heteroaromatic ring conjugation in AMP. The conjugate (a) is the one commonly drawn, and the arrows indicate an electronic rearrangement that leads to another conjugate (b).



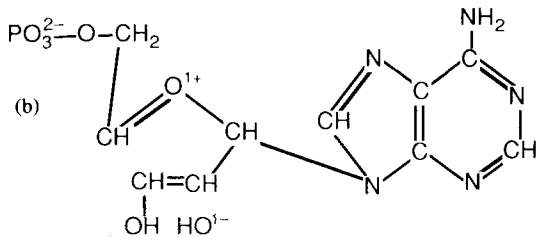


Figure 10. Electronic interactions among groups in AMP due to conjugation. The conjugate (a) is the one commonly drawn, and the arrows indicate an electronic rearrangement that leads to another conjugate (b).

neering Research Center of the University of Maryland; by the Systems Research Center funded by the National Science Foundation (NSFD CDR 8803012); and by a Research Initiation grant from the National Science Foundation (CTS-9010549).

References

1. S.W. Benson, *Thermochemical Kinetics* (Wiley, New York, 1968).
2. S.W. Benson, F. R. Cruickshank, D. M. Golden, G. R. Haugen, H. E. O'Neal, A. S. Rodgers, R. Shaw, and R. Walsh, *Chem. Rev.*, **69,** 279 (1969).
3. E. S. Domalski and E. D. Hearing, *J. Phys. Chem. Ref. Data,* **14,** 1637 (1988).
4. K. G. Joback and R. C. Reid, *Chem. Eng. Commun.*, **57,** 233 (1987).
5. M. L. Mavrovouniotis, P. Bayol, T.-K. M. Lam, G. Stephanopoulos, and G. Stephanopoulos, *Biotechnol. Techn.*, **2,** 23 (1988).
6. R. C. Reid, J. M. Prausnitz, and B. E. Poling, *The Properties of Gases and Liquids,* 4th ed. (McGraw-Hill, New York, 1987).
7. R. C. Reid, J. M. Prausnitz, and T. K. Sherwood, *The Properties of Gases and Liquids,* 3rd ed. (McGraw-Hill, New York, 1977).
8. R. K. Thauer, K. Jungermann, and K. Decker, *Bacteriol. Rev.,* **42,** 148 (1977).
9. T. E. Barman, *Enzyme Handbook, Volume 1* (Springer-Verlag, New York, 1969).
10. T. E. Barman, *Enzyme Handbook, Volume 2* (Springer-Verlag, New York, 1969).
11. T. E. Barman, *Enzyme Handbook* (Springer-Verlag, New York, 1974), Supplement 1.
12. H.-J. Hinz, *Thermodynamic Data for Biochemistry and Biotechnology* (Springer-Verlag, New York, 1986).
13. A. E. Lehninger, *Biochemistry,* 2nd ed. (Worth, New York, 1975).
14. A. E. Lehninger, *Principles of Biochemistry* (Worth, New York, 1986).
15. H. A. Sober, ed., *Handbook of Biochemistry* (CRC, Cleveland, Ohio, 1970).
16. J.T. Edsall and H. Gutfreund, *Biothermodynamics* (Wiley, New York, 1983).
17. R.T. Morrison and R. N. Boyd, *Organic Chemistry,* 3rd ed. (Allyn and Bacon, Boston, 1973).
18. M. L. Mavrovouniotis, "Estimation of Properties from Conjugate Forms of Molecular Structures: The ABC Approach," *Ind. Eng. Chem. Res.*, to appear.
19. M. L. Mavrovouniotis, "Symbolic Computing in the Prediction of Properties of Organic Compounds," Technical Report SRC TR 89-95 (Systems Research Center, University of Maryland, College Park, MD, 1989).