

# MINTA Software Documentation

## Overview

One of the key issues in chemistry is chemical stability. If chemical stability could be reliably predicted by computational methods, then real molecular engineering could be achieved, and one could design stable molecules or molecular complexes having desirable properties rationally. Chemical stability is measured by the free energy of a molecule or molecular complex. The prediction of the relative free energies of different molecular systems is one of the most sought after hopes of computational chemistry. For example, in the pharmaceutical industry chemists often conceive of dozens of molecules they might synthesize but have trouble deciding which ones have the best chance of being potential drug candidates based on their own stability or the stability of molecular complexes they form. Another field of interest is chiral recognition where understanding the stability of molecular complexes involving a chiral reagent can lead to novel chiral resolution techniques. Computational techniques that help the chemists select the most promising candidates for synthesis, or design resolution techniques are extremely valuable. Unfortunately, the thermodynamics of chemical stability is quite complex and use of a state-of-the-art arsenal of computational methods to predict the free energy of molecular systems requires very long computer simulations.

For calculations of the relative binding energies of different ligand molecules for a given receptor to work properly, many different things have to be done correctly. In particular, the gas phase potential energy force field has to be accurate, the effect of solvent has to be included in some realistic and efficient way, and all the vibrational and configurational/conformational states of the system have to be sampled with the correct Boltzmann weights. This last issue is known as the sampling problem and is a particularly difficult

one to solve, because flexible molecules and ligand-receptor complexes may exist in many different conformations. Furthermore, these different conformations may be separated by large energy barriers that prevent these conformations from being inter-converted using traditional simulation methods. An alternative approach embodied in the MINTA software should provide a solution to this problem by affording a direct method for the calculation of conformational and binding free energies without the need for expensive free energy simulations.

The MINTA software incorporates new computational methodology for the direct calculation of free energy without the need for free energy simulations and the application of “computational alchemy”. MINTA utilizes a basic assumption that the total free energy of small to medium sized molecules and molecular complexes is comprised mainly of contributions from the low-energy wells of their respective potential energy surfaces (PES). MINTA relies upon an exhaustive conformational search of the low-energy minima and its basic tenet is a novel multidimensional integration technique that allows, for the first time, for the numerical integration of the high-dimensional configuration integral of individual energy wells. MINTA can, perhaps, be best described in contrast to the well-known quasi-harmonic approximation. Both methods recognize that the local thermodynamics of an energy well can be described by a Boltzmann distribution function. The essence of the quasi-harmonic approximation is to estimate an effective Hessian  $\mathbf{H}$  by calculating the covariance matrix of the internal coordinate variations during a short molecular dynamics (MD) simulation that is local to a particular energy well. The resulting  $\mathbf{H}$  is not equal to the true Hessian  $\mathbf{H}$ , because the effective Hessian includes, implicitly, some anharmonic effects due to the MD simulation. Nevertheless,  $\mathbf{H}$  is used in the context of the harmonic oscillator model to estimate entropy and conformational free energy. MINTA operates exactly the other way around. Instead of sampling the real PES in order to generate an effective Hessian, MINTA utilizes the real Hessian to sample the PES efficiently. The

resulting MINTA integrals of the individual energy wells are, then, summed together to calculate the total molecular configuration integral and the total free energy.

Currently, MINTA can be used for fast computation of the conformational free energy of small and medium sized molecular systems *in vacuo* and in the presence of a continuum solvent model. In particular, MINTA is an excellent tool for the calculation of the binding free energy of molecular complexes comprised of substrate molecules bound to small receptors used in the molecular recognition field or enzyme receptor models often used in pharmaceutical research. Unlike available free energy simulation programs, MINTA calculations are extremely user-friendly and should find wide utility as a simple tool for medicinal chemists already familiar with conformational analysis.

## Introduction

The MINTA methodology is introduced here in the context of calculating binding free energies. The statistical-thermodynamic foundation of the calculation of binding affinities of molecular complexes is quite complex, but for most practical problems, the stability of host-guest complexes can be formulated in terms of binding free energy (BFE) differences. There can be various levels envisioned at which approximations to BFE differences can be made. For example, one wishes to calculate the BFE difference ( $\Delta\Delta G_{L-D} = \Delta G_L - \Delta G_D$ ) between the L and D enantiomers of a ligand bound to an enantioselective host. The simplest approach one can follow is to calculate the energy difference between the lowest energy L and the lowest energy D enantiomer of the ligand. Of course, this approach ignores entropic effects due to the fact that first of all there are multiple binding conformations of both the L and D enantiomers and second of all, the individual conformations are not static (confined to the bottom of their energy well) but exhibit large

dynamic diversity in terms of conformational changes limited to that energy well. Note that there can be numerous low-energy binding conformations of both enantiomeric states.

The next level of approximation to the BFE is the inclusion of multiple conformations. With this model, the multiple low-energy binding conformations of the two enantiomers of the ligand are considered as two sets of discrete energy levels corresponding to the potential energy of the individual binding conformations. A simple statistical mechanics calculation can then be used to estimate the binding free energy difference between the L ligand and the D ligand with respect to the enantioselective host. One can also include some of the vibrational and the rotational free energy utilizing the well-known rigid-rotor harmonic quantum oscillator (RRHO) model. The ultimate approach, in the classical sense, for calculating BFE differences, however, involves the computation of the molecular partition function often termed molecular configuration integral.

For enantioselective binding, for example, the direct calculation of  $\Delta\Delta G_{L-D}$ , again in the classical sense, involves the evaluation of the molecular configuration integral  $Q$ :

$$Q_L = \sum_{i=1}^{n_L} \int_{V_i^L} e^{-\frac{E(\mathbf{r})-E_0}{RT}} d\mathbf{r}, \quad Q_D = \sum_{i=1}^{n_D} \int_{V_i^D} e^{-\frac{E(\mathbf{r})-E_0}{RT}} d\mathbf{r}$$

Equation 1

$$\Delta G_{L-D} = -RT \ln \frac{Q_L}{Q_D}$$

Equation 2

It is assumed with the use of the MINTA software that the dominant part of the configuration integral comes from contributions at or near to low-energy binding conformations. Conformational search results on ligand-receptor complexes suggest that this approximation is feasible for binding free energy calculations. Therefore,  $Q$  is summed over, respec-

tively,  $n_L$  and  $n_D$  conformations each encompassing different  $V$  volumes of the conformational space. Note that the individual terms of the two sums in equation 1 include all the vibrational and configurational states of the particular conformational energy wells of L and D conformations, respectively. Also note that all of the symmetry related copies of a single L or D conformation should be included in the sum in equation 1 to account for the statistical correction for conformational symmetry.  $E(\mathbf{r})$  is the molecular mechanics energy with respect to the nuclear coordinates  $\mathbf{r}$ .  $E(\mathbf{r})$  includes the solvation energy as well, preferably in terms of a continuum model which does not introduce new degrees of freedom by explicit solvent molecules.  $E_0$  is the global minimum energy, which is the common reference for both L and D binding conformations. Thus,  $E_0$  could refer to the lowest energy L or the lowest energy D conformation, whichever is lower.  $R$  is the gas constant and  $T$  is the absolute temperature.

Direct evaluation of the configuration integral has been considered to be impossible to solve except for problems of very low dimensionality. Instead, indirect methods utilizing various simulation techniques based on free energy perturbation (FEP) have found widespread utility. These methods belong to the realm of molecular dynamics and Monte Carlo simulations using explicit solvent models. In the Appendix to this documentation a brief introduction is provided to this area to put MINTA in perspective in the world of free energy simulations.

The basic tenet of the MINTA methodology is a novel Monte Carlo integration technique termed mode integration (this is where the name MINTA comes from). The MINTA software allows, for the first time, the direct calculation of the configuration integral of single molecules and molecular complexes of real chemical interest, without the need for expensive free energy simulations and the use of “computational alchemy”.

## MINTA Methodology

The patented MINTA methodology addresses the sampling problem for calculating free energies at two different levels. First, a global conformational search is carried out to identify the low-energy regions of the potential energy surface (PES), which correspond to the low-energy conformations of a single molecule or the low-energy binding conformations of a molecular complex. Local sampling of each individual conformational energy well is then accomplished utilizing the mode integration technique embodied in the MINTA software. MINTA can, perhaps, be best described in contrast to the well-known quasi-harmonic approximation. Both methods recognize that the local thermodynamics of each energy well ‘*i*’ can be described by a Boltzmann distribution function, which is – in the harmonic approximation – a normalized multivariate Gaussian distribution function:

$$p_i = \sqrt{\frac{\det \mathbf{H}_i}{(2\pi RT)^n}} \exp\left(-\frac{1}{2RT}(\mathbf{r} - \mathbf{r}_i)\mathbf{H}_i(\mathbf{r} - \mathbf{r}_i)\right)$$

Equation 3

where  $\mathbf{r}_i$  and  $\mathbf{H}_i$  denote, respectively, the bottom of a particular energy well and the associated local Hessian matrix. The Hessian is evaluated at the bottom of the well. The number of degrees of freedom  $n$  is equal to the number of unconstrained internal coordinates, which include the six relative translation-rotational degrees of freedom defining the relative orientation of the host and the guest in a molecular complex.  $R$  is the gas constant and  $T$  is the temperature. Note that  $\mathbf{H}_i$ , from the statistical point of view, is none other than the inverse co-variance matrix of the corresponding Gaussian distribution.

The essence of the quasi-harmonic approximation is to estimate an effective Hessian  $\mathbf{H}_i$  by calculating the co-variance matrix of the internal coordinate variations during a short

molecular dynamics (MD) simulation that is local to energy well 'i'. The resulting  $\mathbf{H}_i$  is not equal to  $\mathbf{H}_i$ . The effective Hessian includes, implicitly, some anharmonic effects due to the MD simulation. Nevertheless,  $\mathbf{H}_i$  is used in the context of the harmonic oscillator model to estimate entropy and conformational free energy. MINTA operates exactly the other way around. Instead of sampling the real PES in order to generate an effective Hessian, MINTA utilizes the real Hessian to sample the PES efficiently. However, MINTA sampling is carried out in the context of multidimensional Monte Carlo integration, not some free energy simulation, to afford a unique, direct method for calculating conformational and binding free energies. Mode integration means, in essence, that equation 3 is utilized as a sampling function to integrate equation 1 in *normal mode space*. For further details, the user is referred to the following references:

1. I. Kolossváry; Evaluation of the Molecular Configuration Integral in All Degrees of Freedom for the Direct Calculation of Conformational Free Energies: Prediction of the Anomeric Free Energy of Monosaccharides, *J. Phys. Chem. A* 101, No. 51, 9900-9905 (1997).
2. I. Kolossváry; Evaluation of the Molecular Configuration Integral in All Degrees of Freedom for the Direct calculation of Binding Free Energies: Application to the Enantioselective Binding of Amino Acid Derivatives to Synthetic Host Molecules, *J. Am. Chem. Soc.* 119, 10233-10234 (1997).
3. I. Kolossváry; US Patent Serial No. 08/940, 145, *Mode Integration (MINTA): A Method for Selecting a Molecule Based on Conformational Free Energy of One Molecule for Another Molecule*, filed September 29, 1997.
4. Gy. Keser\_, I. Kolossváry; *Molecular Mechanics and Conformational Analysis in Drug Design, Chapter 7*, Blackwell Science, Oxford, 1999.

## User Manual

The MINTA software is extremely user-friendly. MINTA is launched from BatchMin via a seamless interface that allows the user to invoke MINTA by a single command from a familiar BatchMin command file. The MINTA command has been structured to conform with current BatchMin syntax. A basic sample MINTA command file is shown here:

```
minta.dat
minta.out
FFLD
BGIN
READ
MINI
MNTA
END
```

For users familiar with BatchMin this command structure is well-known, running a so-called multiple minimization job. The input file “minta.dat” should contain one or more conformations of the *same* molecule or molecular complex. The BGIN-END loop reads the structures one by one from the input file and minimizes the energy of the structures with the force field selected in the FFLD command. The MNTA command carries out the MINTA calculation on each and every minimized structure. The output file “minta.out” will contain the minimized structures in the order they were read in.

The arguments of the MNTA command are as follows:

*arg1*                    *Number of MINTA iterations*

The MINTA numerical integrals are calculated in statistical blocks to achieve better convergence. The number of blocks used is referred to as the number of MINTA iterations.

0 5 (default).

*n*                        *n* > 0 (it is *not* recommended to use values *n* > 10).

*arg2*                    *Number of energy evaluations per MINTA iteration*

MINTA integration is based on *single point* energy evaluations. The total number of energy evaluations *per structure* is equal to  $arg1 \times arg2$ .

0 2000 (default).

*n*                         $n > 0$  (it is *not* recommended to use values  $n > 10,000$ ).

*arg3*                    *Flag to choose adaptive MINTA integration*

Adaptive MINTA integration is slightly more accurate than the default, non-adaptive integration mode. *Note however*, that using adaptive MINTA is expected to be beneficial *only* if *arg4* is in the range of 1-10.

0 Non-adaptive MINTA (default).

*n*                         $n \neq 0$  selects adaptive MINTA (*only* recommended when  $1 \leq |arg4| \leq 10$ ).

*arg4*                    *Number of “soft” degrees of freedom, for which numerical MINTA integration is applied*

MINTA integration is carried out in normal mode space. Although MINTA can be instructed to utilize numerical integration in all degrees of freedom, it is far more efficient to partition the degrees of freedom into two categories, “soft” and “hard”, corresponding to low-frequency and high-frequency vibrational modes, respectively. The partition is somewhat arbitrary, of course, but generally “hard” modes can be integrated with high accuracy using an extremely fast analytical approximation to the MINTA integral. “Soft” modes, however, have to be integrated numerically to account for signifi-

cant anharmonic effects. Therefore, a typical MINTA calculation involves numerical integration in “soft-mode” space and analytical integration in “hard-mode” space.

0 Numerical integration in *all* degrees of freedom (default).

*Only recommended for small molecules of up to twenty atoms.*

$n$  For  $|n| > 0$ ,  $|n|$  “soft” modes will be used and the rest of the degrees of freedom will be treated as “hard”.

For  $n > 0$ , the “hard” modes will be integrated utilizing the fast, analytical MINTA approximation.

For  $n < 0$ , the “hard” modes will be integrated utilizing the traditional harmonic oscillator model.

*Note that arg3 should always be zero unless  $1 \leq |n| \leq 10$ .*

*Also note that the user is strongly cautioned not to use values  $|n| > 50$ .*

$arg5$  Temperature (K)

0 300 (default).

$x$   $x > 0$ .

$arg6$  Hard limit for sampling along normal modes ( $\text{\AA}$ )

Sampling will be limited to this distance from the equilibrium geometry of the structure along any of the normal mode directions in  $3N-6(5)$  dimensional normal mode space where  $N$  is the number of atoms.

0 1  $\text{\AA}$  (default).

$x$   $x > 0$  (it is *not* recommended to use values  $x > 3$ ).

$arg7$  Soft limit for sampling along normal modes (units of standard devi-

ation)

Sampling will be limited to different distances from the equilibrium geometry along different normal mode directions. For a particular mode ‘i’ sampling is limited to a particular distance, which is equal to  $arg7$ -times the standard deviation of the multidimensional Gaussian function in equation 3, along the particular normal mode direction ‘i’. The value of distance ‘i’ in Å is  $arg7 \times \sqrt{(RT/\lambda_i)}$  where  $\lambda_i$  is the ‘i’<sup>th</sup> eigenvalue of the Hessian matrix. Note that the softer the mode the larger the sampling distance because of the reciprocal nature of  $\lambda_i$  in the standard deviation formula.

*Important:  $arg6$  takes precedence over  $arg7$ . The actual sampling distance along a particular normal mode ‘i’ will be  $\text{MIN}(arg6, arg7 \times \sqrt{(RT/\lambda_i)})$ . This is an important safeguard to prevent incorrect sampling due to artificially small eigenvalues associated with exceptionally soft, highly anharmonic vibrational modes.*

0                    3 units of standard deviation (default).  
 $x$                      $x > 0$  (it is *not* recommended to use values  $x > 5$ ).

$arg8$                 *Not in use*

DEBUG switch 1001 prints details of the numerical integration in the logfile. The calculated numerical integrals are listed for each statistical block (MINTA iteration) individually and the block averages are also printed along with their  $\chi^2$  test values. Any significant discrepancy of  $\chi^2$  from 1 indicates insufficient sampling and consequently, the MINTA results cannot be trusted. Insufficient sampling can also be detected by looking at the error bar associated with each integral value. The error bar is defined as  $\pm 1$  standard deviation. The debugging output also includes the actual size of the integration box in “soft-mode” space.

## Notes

1. The input file of a MINTA calculation should contain the output of a preceding Batch-Min conformational search.
2. The conformational search should not discard symmetrically equivalent conformations via NSRO, NSRF, ATEQ, and NSEQ commands in order for their statistical weights to be accounted for correctly.
3. The energetic parameters, energetic constraints, and substructure definitions in the MINTA command file (FFLD, SOLV, EXNB, CHGF, FXDI, FXBA, FXTA, SUBS, FXAT) should be identical to those used in the preceding conformational search.
4. MINTA free energy should be looked at exactly the same way as molecular mechanics energy.
  - (i) The MINTA free energy is only meaningful when comparing the difference between two conformations of the same molecule. It estimates the free energy difference between those two conformations.
  - (ii) In the same way, MINTA can be used to calculate the free energy difference between two molecules whose molecular mechanics energy is comparable (typically stereo-isomers of any kind).
  - (iii) For molecules with incomparable molecular mechanics energy, however, MINTA has to be applied in terms of a thermodynamic cycle in order to estimate, for example, the binding free energy difference between two different ligands bound to the same receptor (see Appendix).
  - (iv) **Very important.** The MINTA software calculates the MINTA free energy of each conformation in a multi-conformer input file, but MINTA also calculates the *total free energy* of the whole input file, i.e., the total free energy of the whole set of structures in the input file.

*The total free energy of different multi-conformer files can be compared using the*

*exact same criteria applied to individual conformations.*

For example, the results of a conformational search on glucose are separated in two multi-conformer files,  $\alpha$  anomers in one file and  $\beta$  anomers in a second file. Two MINTA calculations carried out on the two different files will result in the total free energy of the  $\alpha$  anomers and the  $\beta$  anomers, respectively. The difference between the  $\alpha$  and  $\beta$  total free energies provides an estimate for the measurable, so-called anomeric free energy of glucose.

5. As a rule of thumb, the CPU time required to run a nearly complete conformational search on any kind of molecular system is comparable to the CPU time required running a subsequent MINTA calculation.
6. ***Finally, it should be born in mind that MINTA is only as good as the force field and the conformational search.*** MINTA is expected to provide a good estimate of the free energy of a molecular system for a given force field, but MINTA is always subject to serious error due to inadequate force field selection or an incomplete conformational search.

## **Working examples**

### **1. Cyclononane**

A very instructional example is the MINTA calculation on cyclononane. In this example we look at the free energies of the individual conformers of cyclononane. Cyclononane has seven low-energy conformers within 50 kJ/mol above the global minimum on the MM2 potential energy surface. The global minimum is a highly symmetrical ( $C_3$ ), deep minimum, which is 3.14 kJ/mol deeper than the second lowest energy minimum. However, it has been suggested that entropic effects due to the high flexibility of the second and third lowest energy minima could account for decreasing the *free energy gap* between

the global minimum and higher minima. In fact, a simple MINTA calculation suggests that not only a decrease in the free energy gap is predicted, but the (free) energetic order is turned around. *Such (free) energetic reordering of conformations with respect to their energetic order based on steric energy only is an extremely important aspect of molecular design where stable conformations of a molecule are sought.*

In symmetrical molecules, like cyclononane, the statistical weight (due to the existence of symmetry related copies of each conformation) has a significant contribution to the free energy of a particular conformation. There are at least two ways to deal with this. One way is to apply a statistical correction to the free energy that one might calculate for a single copy of each individual conformation. However, with MINTA, we generally recommend using the following (somewhat redundant) alternative:

First, a conformational search is performed, in which the NSRO command is not used (see Note 2, above) **and hence, the symmetry related copies are not discarded.** Then, the single output file resulting from the conformational search is separated into seven individual files, each containing multiple copies of the same conformation.

Thus, in the subdirectory “c9” the following files are provided for running the MINTA calculation on cyclononane:

1. ?.dat where ‘?’ stands for the numbers 1-7. These multi-conformer MacroModel input files contain multiple copies of the seven conformations of cyclononane found by a preceding conformational search. As stated above, the conformational search was run with BatchMin *without* the NSRO command to keep all of the symmetrical copies of the same conformation, subject only to a numbering system rotation around the nine-membered ring. For an asymmetrical conformation this results in nine copies (conformations 2, 3, 6, and 7). The global minimum has  $C_3$  symmetry, therefore, it has only three symmetrical copies due to the numbering system rotation. Conformations 4 and 5, on the other hand, do not have nine but eighteen copies. BatchMin normally

excludes enantiomers from a conformational search, however, conformations 4 and 5 of cyclononane are special in that they are not only equivalent but *identical* to their enantiomers. This is why they both have eighteen copies, not nine. In summary, the statistical weight of the global minimum is only one third of conformations 2, 3, 6, and 7 and one ninth of conformations 4 and 5.

## 2. c9\_minta.com, the MINTA command file:

```
c9_minta.dat
_c9_minta.out
_ DEBG      1001
_ DEBG      601
_ EXNB       0      0      0      0      0.0000      0.0000      0.0000
0.0000
_ FFLD       1      0      0      0      0.0000      0.0000      0.0000
0.0000
BGIN
READ
MINI         9      0      50      0      0.0000      0.0000      0.0000
0.0000
CONV         2      2      0      0      0.0000
MNTA         5    5000      0      50    300.0000      1.0000      3.0000
0.0000
END
```

Note that the individual dat-files have to be renamed c9\_minta.dat to be able to run this particular MINTA command file. DEBG 1001 will print the MINTA debugging output, DEBG 601 prevents MINTA from writing out the minimized structures after minimization. The structures will be written to the output file after the MINTA calculation is completed and their MINTA free energy will be printed in the title line of the structure in c9\_minta.out.

The MINTA calculation runs five iterations with 5,000 energy evaluations in each iteration (25,000 total per conformation). Non-adaptive Monte Carlo integration is applied in “soft-mode” space spanned by the first 50 low-frequency vibrational modes. The rest of the modes are integrated using the fast analytical MINTA method. The temperature is set to 300 K. The size of the integration volume (box) in “soft-mode” space is determined by the lesser of 1 Å and three-times the standard deviation (see *arg6* and *arg7* of the MINTA command description above).

3. ?log where '?' stands for the numbers 1-7. These logfiles are the logfiles of the MINTA calculations 1-7.

Based on the 'Gtotal' values in the logfiles, the MINTA calculation on cyclononane yields the following result:

Order of cyclononane conformations based on (MM2) steric energy:

	E	$\Delta E$ (kJ/mol)	
1.	97.86	0.00	global minimum steric energy
2.	101.00	3.14	
3.	101.10	3.24	
4.	107.16	9.30	
5.	111.10	13.24	
6.	121.59	23.73	
7.	141.13	43.27	

Order of cyclononane conformations based on (MM2) MINTA free energy:

	G	$\Delta G$ (kJ/mol)	
1.(2.)	481.24	0.70	
2.(3.)	481.36	0.82	
3.(1.)	480.54	0.00	global minimum free energy
4.(4.)	484.81	4.27	
5.(5.)	489.32	8.78	
6.(6.)	499.58	19.04	
7.(7.)	521.48	40.94	

## 2. Glucose

In the subdirectory 'glucose' the necessary files are provided for running an anomeric free energy calculation on glucose using the AMBER\* force field and GB/SA continuum solvation for water. In this example, we are looking at the total free energy of each of two sets of conformations (one set for the  $\alpha$  anomer, and one for the  $\beta$  anomer). Note that the  $\alpha$  anomer and  $\beta$  anomer of glucose are diastereomers. Also note that the MINTA calculation on glucose predicts the  $\beta$  anomer to be more stable than the  $\alpha$  anomer by 0.33 kcal/mol, which is within 0.01 kcal/mol of the experimental value (0.34 kcal/mol). However, the calculation is an "overkill", including far too many conformations. The user is encour-

aged to test how many of the approximately 1500 conformations can be omitted to maintain a reasonable accuracy in the calculation.

### 3. Vancomycin

In the subdirectory 'complex' the necessary files are provided for running a binding free energy calculation involving the vancomycin Nac-D-Ala-D-Ala complex using the AMBER\* force field and GB/SA continuum solvation for water.

The 'complex' sample calculation involves fifteen low-energy binding conformations of the vancomycin Nac-D-Ala-D-Ala and Nac-L-Ala-L-Ala complex, respectively. This example has significant relevance to pharmaceutical applications.

Antibiotics of the vancomycin group have gained increasing clinical importance during the last twenty years for the treatment of Gram-positive bacterial infections, particularly those which exhibit multiple drug resistance (Williams, D. H. (1984) *Acc. Chem. Res.* 17, 364; Cohen, M. L. (1992) *Science* 257, 1050; Mackay, J. P.; Gerhard, U.; Beauregard, D. A.; Westwell, M. S.; Searle, M. S.; Williams, D. H. (1994) *J. Am. Chem. Soc.* 116, 4581.). The molecular basis for the mode of action of vancomycin involves the binding of vancomycin to cell-wall mucopeptide precursors terminating in the peptide -D-Ala-D-Ala. The peptide cell-wall analogue NAc-D-Ala-D-Ala has served as a convenient model for X-ray and NMR studies aimed at determining the binding mode of vancomycin. The vancomycin Nac-D-Ala-D-Ala complex represents an interesting "reverse" binding structure. The real binding structure involves the vancomycin substrate binding to the cell-wall whereas the reverse model system involves the cell-wall analogue Nac-D-Ala-D-Ala substrate binding to the vancomycin binding site. Vancomycin has recently been used as the model receptor in several experimental studies for the measurement of binding affinities of peptides to vancomycin (Dunayevskiy, Y. M.; Lyubarskaya, Y. V.; Chou, Y.-H.; Vouros, P.; Karger, B. L. (1998) *J. Med. Chem.* 41, 1201). It has been experimentally determined that while the Nac-D-Ala-D-Ala ligand binds to vancomycin very tightly, the Nac-L-Ala-L-Ala ligand is only a very weak binder. A qualitative/semi-quantitative MINTA calculation is presented in the 'complex' sample where the difference in binding affinity is predicted to be 3.1 kcal/mol in favor of the Nac-D-Ala-D-Ala ligand. Note, however, that a fully quantitative MINTA analysis would require a significantly more complete conformational

search and proof of the aptness of the AMBER\* force field and GB/SA solvation. Again, the user is encouraged to further experiment with vancomycin.

## **Appendix**

Excerpted with permission from Gy. Keser\_, I. Kolossváry; *Molecular Mechanics and Conformational Analysis in Drug Design, Appendix to Chapter 7*. Copyright Blackwell Science, Oxford, 1999.

Ordering information:

<http://www.blackwell-science.com/~cgilib/bookpage.bin?File=5798>

## **Introduction**

In this Appendix we intend to put MINTA in perspective in the world of free energy simulations. Until very recently, MD has been considered to be the only reasonable way of performing free energy simulations (see for example Allen and Tildesley, 1987; McCammon and Harvey, 1987; Straatsma and McCammon, 1992). However, Jorgensen has shown that Metropolis Monte Carlo (MMC) simulation can be equally well utilized for proteins and pointed out that “Monte Carlo sampling of the internal coordinates of protein side chains is likely to be as efficient as molecular dynamics” (Essex et al., 1997). Of course, one has to be very careful with the term Monte Carlo simulation. MMC sampling in the context of free energy perturbation is fundamentally different from MC conformational searching/sampling. The conformational search aspect of MC sampling for protein side chains (Shenkin et al., 1996) and for enzyme active sites (Keser\_ and Kolossváry, 1997) was in fact recognized earlier. Of course, conformational analysis itself does not provide “free energy”, but MINTA does.

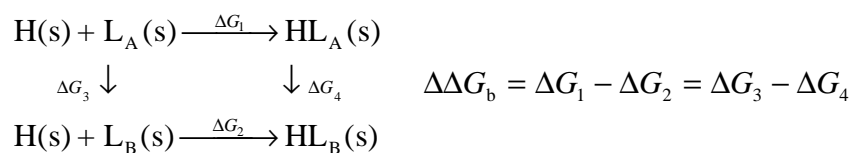
Moreover, it was pointed out by Guida and coworkers (1992) that “presently, thermodynamic perturbation and integration methods seem to be limited to situations in which the

structural perturbation considered does not induce a significant conformational change”, and it was also noted by Kollman (1996) that “the largest challenge facing us is the local minimum, or sampling, problem” and “it is clear that a standard application of molecular dynamics and Monte Carlo methods is very inefficient at traversing the space between local minima”. Free energy simulations using MD or MMC have been applied most successfully to macromolecular complexes where the X-ray structure of the bound complex is known. Although the sampling during MD or MMC simulations allows for the free movement of the ligand, because of the high energetic barriers that impede ligand movement in the active site, MD and MMC tend to keep the ligand close to its bound X-ray conformation, and only the ligand’s internal degrees of freedom are sampled adequately. This approach has been proven, nonetheless, very successful in calculating the binding free energy difference between similar ligands in a relatively rigid active site. However, in a common situation where the bound conformation is not known or the ligand can adopt multiple binding conformations, or when the protein host undergoes significant backbone (e.g. loop) conformational changes upon binding, MD and MMC alone cannot presently provide converged free energy simulation results. Conformational analysis is therefore a prerequisite for any free energy calculation involving highly flexible docking (especially if three-dimensional structural data are not available), whether it is based on MD or MMC simulation, or a direct approach such as MINTA.

Before starting a detailed discussion of free energy simulation techniques and the place for MINTA among them, let us provide recent literature references to the state-of-the-art of calculating protein-ligand binding affinities. Recent reviews include Kollman (1993), Ajay and Murcko (1995), Gilson et al. (1997), Lamb and Jorgensen (1997), and Babine and Bender (1997).

## The thermodynamic cycle

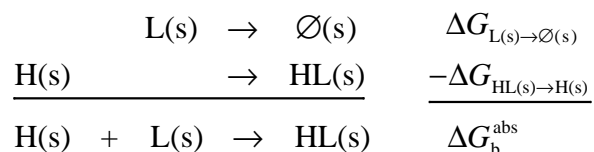
The thermodynamic cycle is based on the fact that free energy is a state function. For example we want to calculate the binding free energy difference between two ligands with respect to the same macromolecular host in solution. The corresponding thermodynamic cycle can be written as follows



Equation A.1

The natural, chemical path would be along the horizontal arrows ( $\Delta G_1 - \Delta G_2$ ) corresponding to the actual chemical complex formation. However, the associated simulations are unfortunately subject to excessive statistical error rendering this approach extremely unstable and virtually useless. On the other hand, the counterintuitive path along the vertical arrows ( $\Delta G_3 - \Delta G_4$ ) is non-chemical, nonetheless, it is well suited for simulations. The computational procedure often referred to as computational alchemy involves “mutating”  $L_a$  into  $L_b$  once in pure solvent and then again in the binding cavity of the protein host H. The numerical values of  $\Delta G_3$  and  $\Delta G_4$  can be obtained by many different ways based on different levels of approximation, which will be discussed below.

Equation A.1 is used to calculate relative binding free energies. The thermodynamic cycle, however, can be used to estimate absolute free energies as well. This procedure is termed double annihilation and the corresponding thermodynamic cycle can be written as a “balance” equation:



Equation A.2

Equation A.2 describes the balance of double annihilation, i.e., in one process the ligand vanishes from the solvent, and in the other process the same ligand vanishes from the ligand-host complex. The net  $\Delta G_{\text{b}}$  is the absolute binding free energy of ligand L to the host H. Note, however that the vanishing of the ligand is intrinsically coupled with the loss of translational and rotational freedom. Hermans and Wang (1997) have recently introduced a restoring potential applied to the ligand that allows for the correct inclusion of the loss of translational and rotational freedom in double annihilation binding free energy calculations.

### Computational methods

The computational methods include a wide palette of approximations. On one end, there are the popular docking algorithms based on empirical scoring functions (to estimate binding free energies) that allow the fast screening of thousands of ligands in the active site of an enzyme model on the time scale of only CPU minutes. Such methods include DOCK (DesJarlais et al., 1988), AutoDock (Goodsell and Olson, 1990; Morris et al., 1996), LUDI (Böhm, 1994), QXP or McDock (McMartin and Bohacek, 1995; McMartin and Bohacek, 1997), Hammerhead (Welch et al., 1996), and SMOG (DeWitte and Shakhnovich, 1996). The primary use of these methods is in the lead finding phase of drug design. Lead finding involves screening of large databases (tens or even hundreds of thousands of small molecules) for potential drug candidates, which hopefully bind to a particular enzyme. The fast screening procedure docks the ligands into the enzyme active site model

and the scoring is based on simple, empirical functions that can reproduce experimental relative binding affinities within 1 to 2 kcal/mol.

On the “high-end” of the list there are highly sophisticated methods based on statistical perturbation theory, in particular free energy perturbation (FEP) and thermodynamic integration (TI). These methods are slow, but represent the highest level of accuracy - ~0.5 kcal/mol for relative binding affinities - that can be achieved today and, therefore, are primarily used for lead optimization. Lead optimization in essence means that small changes are applied to the lead molecules obtained during the lead finding process, trying to enhance their binding affinity by one to three orders of magnitude.

FEP establishes a link between free energy difference  $\Delta G$  and potential energy difference  $\Delta E$  utilizing Zwanzig’s famous equation (Zwanzig, 1954):

$$\Delta G_{AB} = -RT \ln \left\langle e^{-\frac{E_B - E_A}{RT}} \right\rangle_A$$

Equation A.3

$\Delta G_{AB}$  is the free energy difference between two systems “A” and “B”, which can, e.g., represent two different ligands in a situation described by the thermodynamic cycle in Equation A.1.  $E_A$  and  $E_B$  is the potential energy (molecular mechanics energy including solvation energy) of system “A” and “B”, respectively. Note that  $E_A$  and  $E_B$  are functions of the coordinates of the two different systems. The bracket  $\langle \rangle_A$  refers to an ensemble average over system “A”, which is determined at a particular temperature  $T$  ( $R$  is the gas constant).

It should be noted that Equation A.3 is exact. It will be discussed later why is FEP a per-

turbation theory. For the sake of argument, let us assume that we want to use Equation A.3 directly to calculate the binding free energy difference between two ligands  $L_A$  and  $L_B$  with respect to a macromolecular host  $H$ . What does it mean to compute an ensemble average over system “A”? The answer depends on what kind of simulation is applied. If it is MD, the answer is that the MD simulation is running using  $L_A$ ,  $L_A$  is replaced by  $L_B$  after every time step, and the exponential in Equation A.3 is accumulated during the full course of the MD simulation. MD (stochastic dynamics, to be correct) mimics thermal motion in a thermal bath and, therefore, generates the so-called canonical ensemble, i.e., samples the configuration space with the Boltzmann probability. Thus, the simple arithmetic mean of the accumulated exponentials  $\exp(-(E_B-E_A)/RT)$  gives the ensemble average in Equation A.3.

In case of MMC simulation, the Metropolis algorithm guaranties that the canonical ensemble is generated (Metropolis et al., 1953). With MMC, therefore, the well-known Metropolis criterion is applied to  $L_A$  to drive the simulation. Similar to MD, the ensemble average is calculated as the arithmetic mean of the accumulated exponentials  $\exp(-(E_B-E_A)/RT)$  where  $L_A$  is temporarily replaced by  $L_B$  after each accepted MMC move. It is important to note that FEP is always carried out in an explicit solvent box.

Thermodynamic integration provides an alternative way for free energy perturbation and is based on the following formula (Kollman, 1993):

$$\Delta G_{AB} = \int_{\rho=A}^{\rho=B} \left\langle \frac{\partial E_{\rho}}{\partial \rho} \right\rangle_{\rho} d\rho$$

Equation A.4

where the ensemble average  $\langle \rangle_{\rho}$  of the derivative of the energy with respect to  $\rho$  is evaluated at various values of  $\rho$ , and the outer integral is solved numerically.  $\rho$  represents a

“reaction co-ordinate”, which is in most cases non-physical, but delineates a computationally accessible path in phase space to perturb system “A” into system “B”. The ensemble average itself can be calculated the same way as described with FEP.

It is now time to shed light on the perturbation nature of FEP. Although Equation A.3 is exact, evaluation of the ensemble average for systems which differ in more than a trivial way, must be carried out in numerous intermediate stages. The problem is that the “replacement” step in the simulation grossly brakes the thermodynamic equilibrium if, e.g.,  $L_B$  is significantly different from  $L_A$ . Consequently, Equation A.3 will not lead to a sensible free energy. The solution to this problem is making FEP a perturbation procedure by breaking the “mutation” of  $L_A$  into  $L_B$  in several small steps. In other words, one breaks up the free energy calculation into windows, each one involving only a small change whose free energy contribution can be calculated accurately using Equation A.3. Since free energy is a state function, one can add up the small contributions of the subsequent windows to obtain an accurate estimate of the overall  $\Delta G$ :

$$\Delta G_{AB} = \sum_{\lambda=A}^{\lambda=B} -RT \ln \left\langle e^{-\frac{\Delta E_{\lambda}}{RT}} \right\rangle_{\lambda}$$

Equation A.5

where  $\lambda$  is a parameter, which is used to mutate system “A” into system “B” smoothly.  $\Delta E_{\lambda}$  is the potential energy difference between two slightly different, intermediate forms of the hybrid molecule “AB”. Mutation simply means that the geometry and the force field parameters of “A” are continuously changed in small subsequent steps into that of “B”. For example, a  $-\text{CH}_2\text{-OH}$  fragment can be changed to a  $-\text{CH}_2\text{-SH}$  fragment by continuously interpolating the force field parameters between O and S, and adjusting the geometry to adopt longer bond lengths for SH. One can also change the chirality of an atom by continuously interchanging two substituents. The mutation process often requires dummy

atoms that vanish in one molecule, but slowly emerge as real atoms in the other molecule.

In summary, the thermodynamic cycle in Equation A.1 requires two series of FEP simulations following the computational alchemy path. One series will mutate  $L_A$  into  $L_B$  in solvent and the other series will carry out the same mutation in the cavity of the protein. Note that both series consist of several MD or MMC simulations, each of which should sample Boltzmann-weighted ensembles of conformational states adequately to yield converged and accurate free energy values. Therefore, FEP is an extremely time consuming procedure which is always suspect of inadequate sampling (van Gunsteren and Mark, 1992).

A very promising approach termed the linear response method (LRM) (Åqvist et al., 1994; Åqvist, 1996) eliminates the substantial effort devoted to FEP simulations at intermediate points ( $0 < \lambda < 1$ ) along the mutation path. LRM only requires simulations at the end points with the pure systems “A” and “B”. LRM is a semi-empirical method, which is based on the quadratic dependence of free energy on solute charge in the Born model for ion solvation. It can be shown that in this model the electrostatic contribution toward the solvation energy is equal to half of the corresponding ion-solvent interaction energy (Warshel and Russell, 1984; Roux et al., 1990). LRM extends this idea toward protein-ligand systems by linking free energy changes to the interaction energies between solutes (ligands) and their environment (protein). The interactions are broken down into electrostatic and van der Waals contributions. The free energy difference between systems “A” and “B” is given by:

$$\Delta G_{AB} = \frac{1}{2} (\langle E_{\text{ele}} \rangle_A - \langle E_{\text{ele}} \rangle_B) + \alpha (\langle E_{\text{vdw}} \rangle_A - \langle E_{\text{vdw}} \rangle_B)$$

Equation A.6

where  $\alpha$  is an empirical parameter derived from experimental binding data and the ensemble averages  $\langle E_{\text{ele}} \rangle$  and  $\langle E_{\text{vdw}} \rangle$  are obtained via short MD simulations in a periodic box

of water on system “A” and “B”, respectively. LRM has been recently applied successfully to various substrates of cytochrome P450<sub>cam</sub> (Paulsen and Ornstein, 1996) and HIV-1 protease inhibitors (Hultén et al., 1997).

An extended linear response method has been recently introduced (Carlson and Jorgensen, 1995; McDonald et al., 1997) and applied to predicting binding affinities for sulfonamide inhibitors with human thrombin (Jones-Hertzog and Jorgensen, 1997). The extended LRM includes a cavity term and also allows the factor of 0.5 for electrostatic interactions to vary. The resulting free energy equation is

$$\Delta G_{AB} = \beta (\langle E_{\text{ele}} \rangle_A - \langle E_{\text{ele}} \rangle_B) + \alpha (\langle E_{\text{vdw}} \rangle_A - \langle E_{\text{vdw}} \rangle_B) + \gamma (\langle \text{SASA} \rangle_A - \langle \text{SASA} \rangle_B)$$

Equation A.7

where all three empirical parameters ( $\alpha$ ,  $\beta$ ,  $\gamma$ ) are derived from experimental binding data and the third term includes a contribution from the solute’s (ligand) solvent-accessible surface area (SASA).

A novel smart Monte Carlo approach termed jumping-between-wells (JBW) should also be mentioned (Senderowitz et al., 1995; Senderowitz et al., 1997). The JBW method coupled with molecular dynamics involves directly monitoring the populations of various conformations of the two systems “A” and “B” in a simulation in which conformational interconversions occur frequently, producing converged, Boltzmann-weighted ensembles of conformational states. JBW uses a “lookup table” of low-energy conformational states obtained by a preceding conformational search to direct the simulation toward low-energy regions of the PES. Therefore, JBW is not impeded by high barriers like MD, it can simply jump from one minimum to another (subject to the Metropolis criterion) and sample the energy wells locally via MD and MMC. The free energy difference between two systems

can be calculated directly from monitoring the population ratio of the different conformational states of “A” and “B” during the simulation. Note that JBW does not require mutations either, since the simulation is carried out directly on the pure systems.

Direct methods have emerged very recently, which involve the direct evaluation of the configuration integral as sums over conformational minima (Equation 1). Most notably, a new method termed “mining minima” has been introduced in which the configuration integral is evaluated over the “soft modes” identified as torsion angles (Head et al., 1997). It should be stressed, however, that the exclusion of “hard modes” such as bond lengths and bond angles is, in general, a poor approximation. Cyclic structures, e.g., undergo sufficient variation of their ring bond angles and even bond lengths during conformational interconversions, to contribute a significant amount to the conformational free energy. Therefore, a direct method such as MINTA is sought that can evaluate the configuration integral in all degrees of freedom, in order to calculate accurate free energies.

On this palette MINTA can be placed between LRM and FEP. Evaluation of the thermodynamic cycle in Equation A.1 requires four MINTA calculations on  $L_A$ ,  $L_B$ ,  $HL_A$ , and  $HL_B$ , respectively. One MINTA calculation including the conformational search is comparable to a converged MD or MMC simulation on the same system. Therefore, MINTA is intrinsically faster than FEP. On the other hand, MINTA is slower than LRM, because the latter does not require a fully converged simulation. It is too early to make a fair comparison of the accuracy of MINTA vs. FEP or LRM, but based on our results so far, we expect MINTA to be between LRM and FEP in this respect, too. MINTA is certainly subject to two liabilities: (i) the focus on low-energy conformations and (ii) the use of a continuum solvation model. However, the high barriers in a protein-ligand complex with respect to ligand movement in the cavity clearly provide justification for (i) (advantage rather than liability!), and continuum solvation models such as the GBSA model (Still et al., 1990;

Qiu et al., 1997) have recently undergone significant improvements (including the continuum treatment of long-range interactions with application to protein-ligand binding (Simonson et al., 1997)) to render them on par with explicit models for many systems, thus alleviating (ii). In summary, we believe that MINTA should find wide utility as a simple tool for medicinal chemists already familiar with conformational analysis.

### 3. References

- Ajay, Murcko, M. A. (1995) *J. Med. Chem.* 38, 4953.
- Allen, M. P., Tildesley, D. J. (1987) *Computer Simulation of Liquids*, Clarendon Press, Oxford.
- Åqvist, J. (1996) *J. Comput. Chem.* 17, 1587.
- Åqvist, J., Medina, C., Samuelsson, J.-E. (1994) *Protein Engng.* 7, 385.
- Babine, R. E., Bender, S. L. (1997) *Chem. Rev.* 97, 1359.
- Böhm, H.-J. (1994) *J. Comput.-Aided Mol. Design* 8, 243-256.
- Carlson, H. A., Jorgensen, W. L. (1995) *J. Phys. Chem.* 99, 10667.
- DesJarlais, R. L., Sheridan, R. P., Seibel, G. L., Dixon, J. S., Kuntz, I. D., Venkataraghavan, R. (1988) *J. Med. Chem.* 31, 722.
- DeWitte, R. S., Shakhnovich, E. I. (1996) *J. Am. Chem. Soc.* 118, 11733.
- Essex, J. W., Severance, D. L., Jorgensen, W. L. (1997) *J. Phys. Chem. B* 101, 9663.
- Gilson, M. K., Given, J. A., Bush, B. L., McCammon, J. A. (1997) *Biophys. J.* 72, 1047.
- Goodsell, D. S., Olson, A. J. (1990) *Proteins* 8, 195.
- Guida, W. C., Bohacek, R. S., Erion, M. D. (1992) *J. Comput. Chem.* 13, 214.
- Head, M. S., Given, J. A., Gilson, M. K. (1997) *J. Phys. Chem. A* 101, 1609.
- Hermans, J., Wang, L. (1997) *J. Am. Chem. Soc.* 119, 2707.
- Hultén, J., Bonham, N. M., Nillroth, U., Hansson, T., Zuccarello, G., Bouzide, A., Åqvist, J., Classon, B., Danielson, U. H., Karlén, A., Kvarnström, I., Samuelsson, B., Hall-

- berg, A. (1997) *J. Med. Chem.* 40, 885.
- Jones-Hertzog, D. K., Jorgensen, W. L. (1997) *J. Med. Chem.* 40, 1539.
- Jorgensen, W. L. (1989) *Acc. Chem. Res.* 22, 184.
- Keser\_, G. M., Kolossváry, I., Bertók, B. (1997) *J. Am. Chem. Soc.* 119, 5126.
- Kollman, P. A. (1993) *Chem. Rev.* 93, 2395.
- Kollman, P. A. (1996) *Acc. Chem. Res.* 29, 461.
- Lamb, M. L., Jorgensen, W. L. (1997) *Curr. Op. Chem. Biol.* 1, 449.
- McCammon, J. A., Harvey, S. C. (1987) *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge.
- McDonald, N. A., Carlson, H. A., Jorgensen W. L. (1997) *J. Phys. Org. Chem.* 10, 563.
- McMartin, C., Bohacek, R. S. (1995) *J. Comput.-Aided Mol. Design* 9, 237.
- McMartin, C., Bohacek, R. S. (1997) *J. Comput.-Aided Mol. Design* 11, 333.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A., Teller E. (1953) *J. Chem. Phys.* 21,1087.
- Morris, G. M., Goodsell, D. S., Huey, R., Olson, A. J. (1996) *J. Comput.-Aided Mol. Design* 10, 293.
- Paulsen, M. D., Ornstein, R. L. (1996) *Protein Engng.* 9, 567.
- Qiu, D., Shenkin, P. S., Hollinger, F. P., Still, W. C. (1997) *J. Phys. Chem. A* 101, 3005.
- Roux, B., Yu, H.-A., Karplus, M. (1990) *J. Phys. Chem.* 94, 4683.
- Senderowitz, H., Guarnieri, F., Still, W. C. (1995) *J. Am. Chem. Soc.* 117, 8211.
- Senderowitz, H., McDonald, D. Q., Still, W. C. (1997) *J. Org. Chem.* 62, 9123.
- Shenkin, P. S., Farid, H., Fetrow J. S. (1996) *Proteins, Struct. Funct. Genet.* 26, 323.
- Simonson, T., Archontis, G., Karplus, M. (1997) *J. Phys. Chem. B* 101, 8349.
- Still, W. C., Tempczyk, A., Hawley, R. C., Hendrickson, T. (1990) *J. Am. Chem. Soc.* 112, 6127.
- Straatsma, T. P., McCammon, J. A. (1992) *Ann. Rev. Phys. Chem.* 43, 407.
- Van Gunsteren, W. F., Mark, A. E. (1992) *Eur. J. Biochem.* 204, 947.

Warshel, A., Russell, S. T. (1984) Q. Rev. Biophys. 17, 283.

Welch, W., Ruppert, J., Ajay, N. J. (1996) Chemistry & Biology 3, 449.

Zwanzig, R. W. (1954) J. Chem. Phys. 22, 1420.