Protein abundances vs. equilibrium activities

Jeffrey M. Dick

August 17, 2011

1 Introduction

The `diagram()` function serves multiple purposes that might be confusing to the new user. From its name, we know that it produces diagrams of some sort. These are equilibrium chemical activity diagrams – that is the primary purpose of the function. However, inspecting the arguments to the function reveals that the input values are the affinities of formation reactions of species in the system. How do we go from chemical affinities to chemical activities? This problem defines the purpose of two auxiliary functions (`equil.react()` and `equil.boltz()`) whose algorithms are described below.

Some explanation of terminology is in order. By chemical activity we mean the quantity a_i that appears in the expression

$$\mu_i = \mu_i^\circ + RT \ln a_i \, , \tag{1}$$

where μ_i and μ_i° stand for the chemical potential and the standard chemical potential of the ith species, and R and T represent the gas constant and the temperature in Kelvin. Chemical activity is related to molality (m_i) by

$$a_i = \gamma_i m_i \, , \tag{2}$$

where γ_i stands for the activity coefficient of the ith species. For this discussion, we take $\gamma_i = 1$ for all species, so chemical activity is assumed to be numerically equivalent to molality. Since molality is a measure of concentration, calculating the equilibrium chemical activities can be a theoretical tool to help understand the relative abundances of species, including proteins.

After going over the methods used in CHNOSZ for equilibrium activity calculations, some comparisons with experimental protein abundance data are made.

2 Calculations at a single point

Here we discuss two procedures for calculating equilibrium activities of species. The first is a reaction-matrix approach and the second takes advantage of the Boltzmann distribution. We show (by example) that the two approaches are equivalent when the formation reactions of residue equivalents of proteins are used. The example system here has also been described in a paper [3].

2.1 Reaction-matrix approach

The next two sections give examples of calculating the equilibrium activities of two proteins using a matrix of equations representing reactions to form the proteins. Although the examples below include only two proteins, each additional protein introduces one more equation and unknown, so this procedure can be carried out for any number of proteins given the necessary computational requirements.

2.1.1 Whole proteins

Let us calculate the equilibrium activities of two proteins in metastable equilibrium. To do this we start by writing the formation reactions of each protein as

$$stuff_3 \rightleftharpoons \mathrm{CSG\_METVO} \tag{3}$$

and

$$stuff_4 \rightleftharpoons \mathrm{CSG\_METJA}\,. \tag{4}$$

The basis species in the reactions are collectively symbolized by $stuff$; the subscripts simply refer to the reaction number in this document. In these examples, $stuff$ consists of CO_2, H_2O, NH_3, O_2, H_2S and H^+ in different molar proportions. To see what $stuff$ is, try out these commands in CHNOSZ:

```
> library(CHNOSZ)
> basis("CHNOS+")

    C H N O S Z ispecies logact state
CO2 1 0 0 2 0 0       69     -3    aq
H2O 0 2 0 1 0 0        1      0   liq
NH3 0 3 1 0 0 0       68     -4    aq
H2S 0 2 0 0 1 0       70     -7    aq
O2  0 0 0 2 0 0     2691    -80   gas
H+  0 1 0 0 0 1        3     -7    aq

> species("CSG", c("METVO", "METJA"))

protein: found CSG_METVO (C2575H4097N645O884S11, 553 residues)
protein: found CSG_METJA (C2555H4032N640O865S14, 530 residues)
   CO2  H2O NH3 H2S      O2 H+ ispecies logact state      name
1 2575 1070 645  11 -2668.0  0     2926     -3    aq CSG_METVO
2 2555 1042 640  14 -2643.5  0     2927     -3    aq CSG_METJA
```

Although the basis species are defined, the temperature is not yet specified, so it is not immediately possible to calculate the ionization states of the proteins. That is why the coefficient on H^+ is zero in the output above. To see what the computed protein charges are at 25 °C and 1 bar and at pH 7 (which is the opposite of the logarithm of activity of H^+ in the basis species), try this:

```
> protein.info()

affinity: temperature is 25 C
energy.args: pressure is Psat
affinity: loading ionizable protein groups
subcrt: 25 species at 298.15 K and 1 bar (wet)
affinity: temperature is 25 C
energy.args: pressure is Psat
subcrt: 25 species at 298.15 K and 1 bar (wet)
info: 2926 refers to CSG_METVO, C2575H4097N645O884S11 aq (BBA+03)
info: 2927 refers to CSG_METJA, C2555H4032N640O865S14 aq (BBA+03)
   CO2  H2O NH3 H2S      O2 H+ ispecies logact state      name
1 2575 1070 645  11 -2668.0  0     2926     -3    aq CSG_METVO
2 2555 1042 640  14 -2643.5  0     2927     -3    aq CSG_METJA
protein.info: converting things ...
    protein length               formula         G      Z       G.Z     ZC
1 CSG_METVO    553 C2575H4097N645O884S11 -24880.93 -56.07 -24976.76 -0.144
2 CSG_METJA    530 C2555H4032N640O865S14 -24236.26 -55.87 -24413.72 -0.139
```

Note that `affinity()` is called twice by `protein.info()`; this so that both charges and standard Gibbs energies of ionization of the proteins can be calculated. The `Z` values in the table are the charges of the proteins computed using the ionization constants of sidechain and terminal groups, and the `G.Z` values are the calculated Gibbs energies of formation of the ionized proteins [5]. The `ZC` values are the average oxidation states of carbon of the proteins. Let us now calculate the chemical affinities of formation of the ionized proteins:

```
> a <- affinity()

affinity: temperature is 25 C
energy.args: pressure is Psat
affinity: loading ionizable protein groups
subcrt: 25 species at 298.15 K and 1 bar (wet)

> a$values

$`2926`
[1] 107.6774

$`2927`
[1] 317.1877
```

Since `affinity()` returns a list with a lot of information (such as the basis species and species definitions) the last command was written to only print the `values` part of that list. The values are actually dimensionless, i.e. $\boldsymbol{A}/2.303RT$.

The affinities of the formation reactions above were calculated for a *reference value of activity of the proteins, which is not the equilibrium value.* Those non-equilibrium activities were 10^{-3}. How do we calculate the equilibrium values? Let us write specific statements of the expression for chemical affinity (2.303 is used here to stand for $\ln 10$),

$$\boldsymbol{A} = 2.303RT\log(K/Q)\,, \tag{5}$$

for Reactions 3 and 4 as

$$\begin{aligned}\boldsymbol{A}_3/2.303RT &= \log K_3 - \log Q_3\\ &= \log K_3 + \log a_{stuff,3} - \log a_{\mathrm{CSG\_METVO}}\\ &= \boldsymbol{A}_3^*/2.303RT - \log a_{\mathrm{CSG\_METVO}}\end{aligned} \tag{6}$$

and

$$\begin{aligned}\boldsymbol{A}_4/2.303RT &= \log K_4 - \log Q_4\\ &= \log K_4 + \log a_{stuff,4} - \log a_{\mathrm{CSG\_METJA}}\\ &= \boldsymbol{A}_4^*/2.303RT - \log a_{\mathrm{CSG\_METJA}}\,.\end{aligned} \tag{7}$$

The $\boldsymbol{A}^*$ denote the affinities of the formation reactions when the activities of the proteins are zero. From the output above it follows that $\boldsymbol{A}_3^*/2.303RT = 104.6774$ and $\boldsymbol{A}_4^*/2.303RT = 314.1877$.

Next we must specify how reactions are balanced in this system: what is conserved during transformations between species (let us call it the immobile component)? For proteins, one possibility is to use the repeating protein backbone group. Let us use n_i to designate the number of residues in the ith protein, which is equal to the number of backbone groups, which is equal to the length of the sequence. If $\gamma_i = 1$ in Eq. (2), the relationship between the activity of the ith protein (a_i) and the activity of the residue equivalent of the ith protein ($a_{residue,i}$) is

$$a_{residue,i} = n_i \times a_i\,. \tag{8}$$

We can use this to write a statement of mass balance:

$$553 \times a_{\mathrm{CSG\_METVO}} + 530 \times a_{\mathrm{CSG\_METJA}} = 1.083\,. \tag{9}$$

At equilibrium, the affinities of the formation reactions, per conserved quantity (in this case protein backbone groups) are equal. Therefore $\boldsymbol{A} = \boldsymbol{A}_3/553 = \boldsymbol{A}_4/530$ is a condition for equilibrium. Combining this with Eqs. (6) and (7) gives

$$\boldsymbol{A}/2.303RT = \left(104.6774 - \log a_{\mathrm{CSG\_METVO}}\right)/553 \tag{10}$$

and

$$\boldsymbol{A}/2.303RT = \left(314.1877 - \log a_{\mathrm{CSG\_METJA}}\right)/530\,. \tag{11}$$

Now we have three equations (9–11) with three unknowns. The solution can be displayed in CHNOSZ as follows. The argument `residue=FALSE` overrides the default setting for `diagram` when proteins are the species of interest and instructs it to use the function named `equil.react()`, which implements the equation-solving strategy described in the next section. Here we retrieve the equilibrium activities using `diagram()` without letting it actually do any plotting.

```
> d <- diagram(a, residue = FALSE, do.plot = FALSE)

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 553 530
diagram: log total activity of PBB (from species) is 0.03462846

> d$logact

$`2926`
[1] -225.9512

$`2927`
[1] -2.689647
```

Those are the logarithms of the equilibrium activities of the proteins. Combining these values with either Eqs. (10) or (11) gives us the same value for affinity of the formation reactions per residue (or per protein backbone group), $\boldsymbol{A}/2.303RT = 0.5978817$. Equilibrium activities that differ by such great magnitudes make it appear that the proteins are very unlikely to coexist in metastable equilibrium. Later we explain the concept of using residue equivalents of the proteins to achieve a different result.

2.1.2 Implementing the reaction-matrix approach

The implementation used in CHNOSZ for finding a solution to the system of equations relies on a difference function for the activity of the immobile component. The steps to obtain this difference function are:

1. Set the total activity of the immobile (conserved) component as a_{ic} (e.g., the 1.083 in Eqn. 9).
2. Write a function for the logarithm of activity of each of the species of interest: $\boldsymbol{A} = \left(\boldsymbol{A}_i^* - 2.303RT\log a_i\right)/n_{\mathrm{ic},i}$, where $n_{\mathrm{ic},i}$ stands for the number of moles of the immobile component that react in the formation of one mole of the ith species. (e.g., for systems of proteins where the backbone group is conserved, $n_{\mathrm{ic},i}$ is the same as n_i in Eq. 8). Calculate values for each of the $\boldsymbol{A}_i^*$. Metastable equilibrium is implied by the identity of $\boldsymbol{A}$ in all of the equations.
3. Write a function for the total activity of the immobile component: $a_{\mathrm{ic}}^{'} = \sum n_{\mathrm{ic},i}a_i$.
4. The difference function is now $\delta a_{\mathrm{ic}} = a_{\mathrm{ic}}^{'} - a_{\mathrm{ic}}$.

Now all we have to do is solve for the value of $\boldsymbol{A}$ where $\delta a_{\mathrm{ic}} = 0$. This is achieved in the code by first looking for a range of values of $\boldsymbol{A}$ where at one end $\delta a_{\mathrm{ic}} < 0$ and at the other end $\delta a_{\mathrm{ic}} > 0$, then using the `uniroot()` function that is part of R to find the solution.

This approach is subject to failure if for all trial ranges of $\boldsymbol{A}$ the δa_{ic} are of the same sign, which gives an error message like "i tried it 1000 times but can't make it work". Even if values of δa_{ic} on either side of zero can be located, the algorithm does not guarantee an accurate solution and may give a warning about poor convergence if a certain (currently hard-coded) tolerance is not reached.

2.1.3 Residue equivalents

Let us consider the formation reactions of the *residue equivalents* of proteins, for example

$$stuff_{12} \rightleftharpoons \text{CSG\_METVO(residue)} \tag{12}$$

and

$$stuff_{13} \rightleftharpoons \text{CSG\_METJA(residue)}\,. \tag{13}$$

The formulas of the residue equivalents are those of the proteins divided by the number of residues in each protein. With the `residue.info()` function it is possible to see the coefficients on the basis species in these reactions:

```
> residue.info()

affinity: temperature is 25 C
energy.args: pressure is Psat
affinity: loading ionizable protein groups
subcrt: 25 species at 298.15 K and 1 bar (wet)
       CO2      H2O      NH3        H2S        O2          H+      name
1 4.656420 1.934901 1.166365 0.01989150 -4.824593 -0.1013835 CSG_METVO
2 4.820755 1.966038 1.207547 0.02641509 -4.987736 -0.1054156 CSG_METJA
```

Let us denote by $\boldsymbol{A}_{12}$ and $\boldsymbol{A}_{13}$ the chemical affinities of Reactions 12 and 13. We can write

$$\boldsymbol{A}_{12}/2.303RT = \log K_{12} + \log a_{stuff,12} - \log a_{\mathrm{CSG\_METVO(residue)}} \tag{14}$$

and

$$\boldsymbol{A}_{13}/2.303RT = \log K_{13} + \log a_{stuff,13} - \log a_{\mathrm{CSG\_METJA(residue)}}\,, \tag{15}$$

For metastable equilibrium we have $\boldsymbol{A}_{12}/1 = \boldsymbol{A}_{13}/1$. The 1's in the denominators are there as a reminder that we are still conserving residues, and that each reaction now is written for the formation of a single residue equivalent. So, let us write $\boldsymbol{A}$ for $\boldsymbol{A}_{12} = \boldsymbol{A}_{13}$ and also define $\boldsymbol{A}^*_{12} = \boldsymbol{A}_{12} + 2.303RT \log a_{\mathrm{CSG\_METVO(residue)}}$ and $\boldsymbol{A}^*_{13} = \boldsymbol{A}_{13} + 2.303RT \log a_{\mathrm{CSG\_METJA(residue)}}$. At the same temperature, pressure and activities of basis species and proteins as shown in the previous section, we can write $\boldsymbol{A}^*_{12} = \boldsymbol{A}^*_3/553 = 2.303RT \times 0.1892901$ and $\boldsymbol{A}^*_{13} = \boldsymbol{A}^*_4/530 = 2.303RT \times 0.5928069$ to give

$$\boldsymbol{A}/2.303RT = 0.1892901 - \log a_{\mathrm{CSG\_METVO(residue)}} \tag{16}$$

and

$$\boldsymbol{A}/2.303RT = 0.5928069 - \log a_{\mathrm{CSG\_METJA(residue)}}\,, \tag{17}$$

which are equivalent to Equations 12 and 13 in the paper [3] but with more decimal places shown. A third equation arises from the conservation of amino acid residues:

$$a_{\mathrm{CSG\_METVO(residue)}} + a_{\mathrm{CSG\_METJA(residue)}} = 1.083\,. \tag{18}$$

The solution to these equations is $a_{\mathrm{CSG\_METVO(residue)}} = 0.3065982$, $a_{\mathrm{CSG\_METJA(residue)}} = 0.7764018$ and $\boldsymbol{A}/2.303RT = 0.7027204$.

The corresponding logarithms of activities of the proteins are $\log(0.307/553) = -3.256$ and $\log(0.776/530) = -2.834$. These activities of the proteins are much closer to each other than those calculated using formation reactions for whole protein formulas, so this result seems more compatible with the actual coexistence of proteins in nature.

The approach just described is not used by `diagram()` when `residue=TRUE` (which is the default setting). Instead, the Boltzmann distribution, described next, is implemented for that situation.

2.2 Boltzmann distribution

An expression for Boltzmann distribution, relating equilibrium activities of species to the affinities of their formation reactions, can be written as (using the same definitions of the symbols above)

$$\frac{a_i}{\sum a_i} = \frac{e^{\boldsymbol{A}^*_i/RT}}{\sum e^{\boldsymbol{A}^*_i/RT}}\,. \tag{19}$$

Using this equation, we can very quickly (without setting up a system of equations) calculate the equilibrium activities of proteins using their residue equivalents. Above, we saw $\boldsymbol{A}^*_{12}/2.303RT = 0.1892901$ and

$\boldsymbol{A}^*_{13}/2.303RT = 0.5928069$. Multiplying by $\ln 10 = 2.302585$ gives $\boldsymbol{A}^*_{12}/RT = 0.4358565$ and $\boldsymbol{A}^*_{13}/RT = 1.364988$. We then have $e^{\boldsymbol{A}^*_{12}/RT} = 1.546287$ and $e^{\boldsymbol{A}_{13}/RT} = 3.915678$. This gives us $\sum e^{A^*_i/RT} = 5.461965$, $a_{12}/\sum a_i = 0.2831009$ and $a_{13}/\sum a_i = 0.7168991$. Since $\sum a_i = 1.083$, we arrive at $a_{12} = 0.3065982$ and $a_{13} = 0.7764018$, the same result as above. This example was also described in a recent paper [4].

This computation can be carried out in CHNOSZ using the following commands, which implies `residue=TRUE` as the default setting for systems of proteins. This setting signifies to consider the formation reactions of the residue equivalents instead of the whole proteins, AND consequently to make a call to `equil.boltz()` rather than `equil.react()`.

```
> d <- diagram(a, do.plot = FALSE)

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 553 530
diagram: using residue equivalents
diagram: log total activity of PBB (from species) is 0.03462846

> as.numeric(d$logact)

[1] -3.256155 -2.834189
```

We can also specify `as.residue=TRUE` (which means to return the logarithms of activities of the residue equivalents rather than converting them to logarithms of activities of the proteins):

```
> d <- diagram(a, as.residue = TRUE, do.plot = FALSE)

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 553 530
diagram: using residue equivalents
diagram: log total activity of PBB (from species) is 0.03462846

> 10^as.numeric(d$logact)

[1] 0.3065982 0.7764018
```

Although this example includes only two proteins, this procedure is suitable for calculating the metastable equilibrium activities of any number of proteins.

3 Calculations as a function of a single variable

A comparison of the outcomes of equilibrium calculations that do and do not use the residue equivalents for proteins was given in a publication [3]. An expanded version of a diagram in that paper is below (though, without labels on the figures).

```
> organisms <- c("METSC", "METJA", "METFE", "HALJP", "METVO", "METBU",
+     "ACEKI", "BACST", "BACLI", "AERSA")
> proteins <- c(rep("CSG", 6), rep("SLAP", 4))
> basis("CHNOS+")

    C H N O S Z ispecies logact state
CO2 1 0 0 2 0 0       69     -3    aq
H2O 0 2 0 1 0 0        1      0   liq
NH3 0 3 1 0 0 0       68     -4    aq
H2S 0 2 0 0 1 0       70     -7    aq
O2  0 0 0 2 0 0     2691    -80   gas
H+  0 1 0 0 0 1        3     -7    aq

> species(proteins, organisms)
```

```
protein: found CSG_METSC (C2812H4405N747O872S16, 571 residues)
protein: found CSG_METFE (C2815H4411N747O872S14, 571 residues)
protein: found CSG_HALJP (C3669H5647N971O1488, 828 residues)
protein: found CSG_METBU (C1362H2111N355O442S4, 278 residues)
protein: found SLAP_ACEKI (C3584H5648N926O1138S4, 736 residues)
protein: found SLAP_BACST (C5676H9113N1489O1863S3, 1198 residues)
protein: found SLAP_BACLI (C3977H6396N1068O1286S2, 844 residues)
protein: found SLAP_AERSA (C2250H3580N618O716S2, 481 residues)
    CO2  H2O  NH3 H2S      O2 H+ ispecies logact state       name
1  2812 1066  747  16 -2909.0  0     2928     -3    aq  CSG_METSC
2  2555 1042  640  14 -2643.5  0     2927     -3    aq  CSG_METJA
3  2815 1071  747  14 -2914.5  0     2929     -3    aq  CSG_METFE
4  3669 1367  971   0 -3608.5  0     2930     -3    aq  CSG_HALJP
5  2575 1070  645  11 -2668.0  0     2926     -3    aq  CSG_METVO
6  1362  519  355   4 -1400.5  0     2931     -3    aq  CSG_METBU
7  3584 1431  926   4 -3730.5  0     2932     -3    aq SLAP_ACEKI
8  5676 2320 1489   3 -5904.5  0     2933     -3    aq SLAP_BACST
9  3977 1594 1068   2 -4131.0  0     2934     -3    aq SLAP_BACLI
10 2250  861  618   2 -2322.5  0     2935     -3    aq SLAP_AERSA
```

```
> a <- affinity(O2 = c(-100, -65))
```

```
affinity: temperature is 25 C
energy.args: pressure is Psat
energy.args: variable 1 is O2 at 128 increments from -100 to -65
affinity: loading ionizable protein groups
subcrt: 33 species at 298.15 K and 1 bar (wet)
```

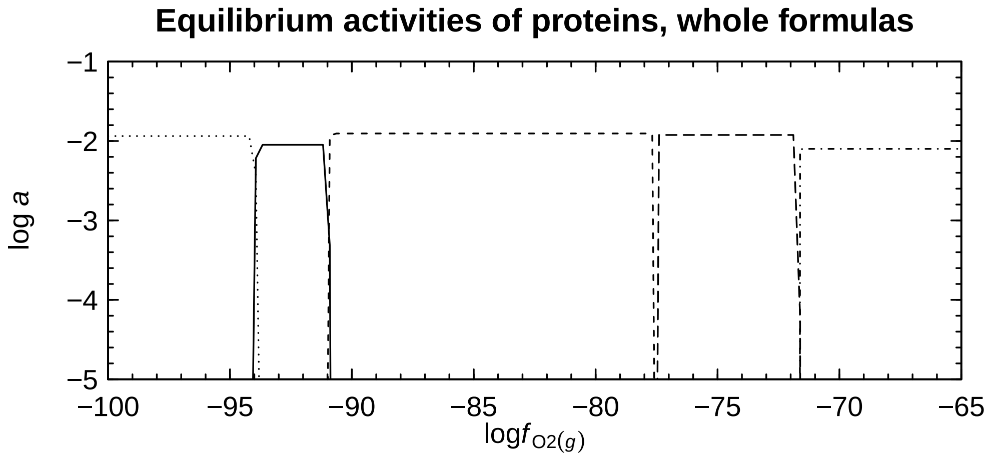
```
> par(mfrow = c(2, 1))
> diagram(a, ylim = c(-5, -1), legend.x = NULL, residue = FALSE)
```

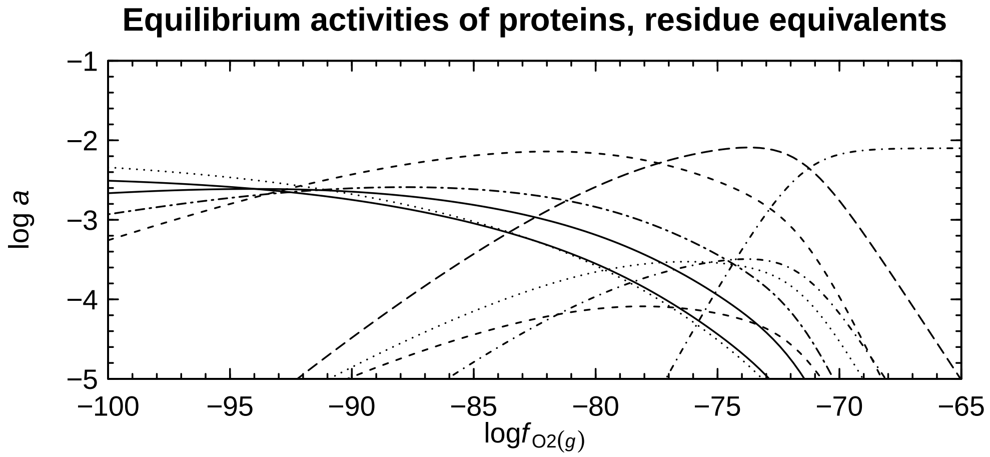
```
diagram: immobile component is protein backbone group
diagram: conservation coefficients are 571 530 571 828 553 278 736 1198 844 481
diagram: log total activity of PBB (from species) is 0.8188854
diagram: poor convergence in step 34 (remainder in logact of -0.00103380130884601)
diagram: poor convergence in step 104 (remainder in logact of -0.00170844472820764)
```

```
> title(main = "Equilibrium activities of proteins, whole formulas")
> diagram(a, ylim = c(-5, -1), legend.x = NULL)
```

```
diagram: immobile component is protein backbone group
diagram: conservation coefficients are 571 530 571 828 553 278 736 1198 844 481
diagram: using residue equivalents
diagram: log total activity of PBB (from species) is 0.8188854
```

```
> title(main = "Equilibrium activities of proteins, residue equivalents")
```

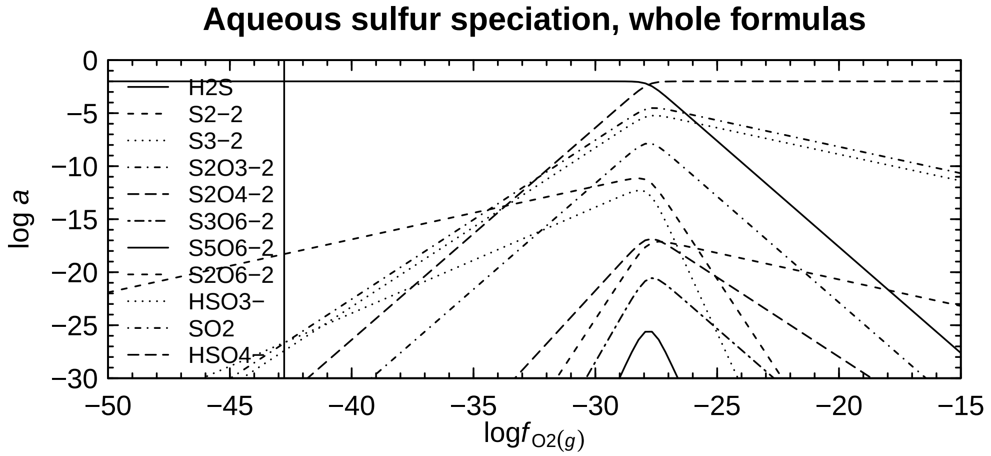


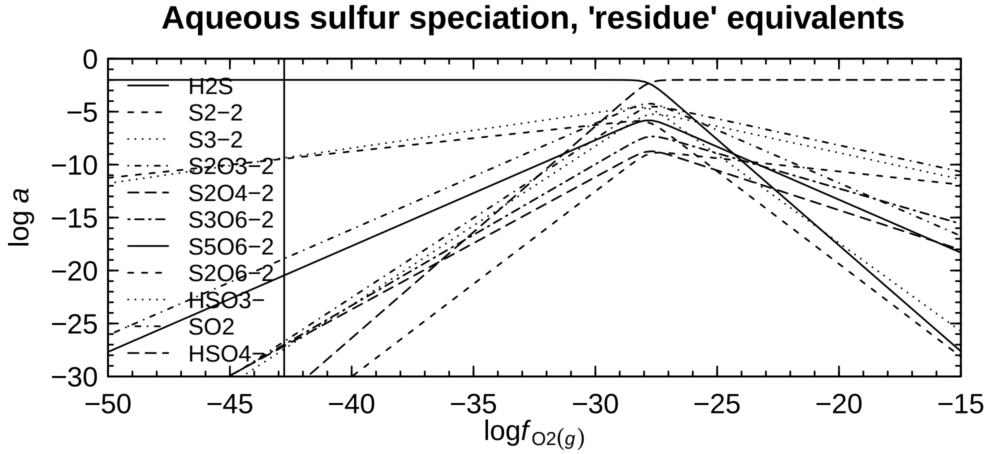


The reaction-matrix approach described above can also be applied to systems having conservation coefficients that differ from unity, such as many mineral and inorganic systems, where the immobile component has different molar coefficients in the formulas. For example, consider a system like that described by Seewald, 1997 [7]:

```
> basis("CHNOS+")

    C H N O S Z ispecies logact state
CO2 1 0 0 2 0 0       69     -3    aq
H2O 0 2 0 1 0 0        1      0   liq
NH3 0 3 1 0 0 0       68     -4    aq
H2S 0 2 0 0 1 0       70     -7    aq
O2  0 0 0 2 0 0     2691    -80   gas
H+  0 1 0 0 0 1        3     -7    aq

> basis("pH", 5)

    C H N O S Z ispecies logact state
CO2 1 0 0 2 0 0       69     -3    aq
H2O 0 2 0 1 0 0        1      0   liq
NH3 0 3 1 0 0 0       68     -4    aq
H2S 0 2 0 0 1 0       70     -7    aq
O2  0 0 0 2 0 0     2691    -80   gas
H+  0 1 0 0 0 1        3     -5    aq
```

```
> species(c("H2S", "S2-2", "S3-2", "S2O3-2", "S2O4-2", "S3O6-2",
+     "S5O6-2", "S2O6-2", "HSO3-", "SO2", "HSO4-"))

   CO2 H2O NH3 H2S  O2 H+ ispecies logact state   name
1    0   0   0   1 0.0  0       70     -3    aq    H2S
2    0  -1   0   2 0.5 -2       53     -3    aq   S2-2
3    0  -2   0   3 1.0 -2       54     -3    aq   S3-2
4    0  -1   0   2 2.0 -2       26     -3    aq S2O3-2
5    0  -1   0   2 2.5 -2     1072     -3    aq S2O4-2
6    0  -2   0   3 4.0 -2     1077     -3    aq S3O6-2
7    0  -4   0   5 5.0 -2     1079     -3    aq S5O6-2
8    0  -1   0   2 3.5 -2     1076     -3    aq S2O6-2
9    0   0   0   1 1.5 -1       23     -3    aq  HSO3-
10   0  -1   0   1 1.5  0       78     -3    aq    SO2
11   0   0   0   1 2.0 -1       25     -3    aq  HSO4-

> a <- affinity(O2 = c(-50, -15), T = 325, P = 350)

affinity: temperature is 325 C
affinity: pressure is 350 bar
energy.args: variable 1 is O2 at 128 increments from -50 to -15
subcrt: 17 species at 598.15 K and 350 bar (wet)

> par(mfrow = c(2, 1))
> diagram(a, logact = -2, ylim = c(-30, 0), legend.x = "topleft",
+     cex.names = 0.8)

diagram: immobile component is H2S
diagram: conservation coefficients are 1 2 3 2 2 3 5 2 1 1 1
diagram: log total activity of H2S (from argument) is -2

> title(main = "Aqueous sulfur speciation, whole formulas")
> diagram(a, logact = -2, ylim = c(-30, 0), legend.x = "topleft",
+     cex.names = 0.8, residue = TRUE)

diagram: immobile component is H2S
diagram: conservation coefficients are 1 2 3 2 2 3 5 2 1 1 1
diagram: using residue equivalents
diagram: log total activity of H2S (from argument) is -2

> title(main = "Aqueous sulfur speciation, 'residue' equivalents")
```





The first diagram is quantitatively similar to the one shown by Seewald, 1997, but in the second (where we have set `residue=TRUE`) the range of activities is lower at any given $\log f_{O_{2(g)}}$. There, the function was told to rewrite the formation reactions of the aqueous sulfur species for their residue equivalents in the same way the formation reactions for the proteins were rewritten above. The number of "residues" in each species is the coefficient of the immobile component, in this case H_2S, in the formation reaction.

Maybe `residue=TRUE` doesn't make sense for systems where the formulas of species are similar in size to those of the basis species. For molecules as large as proteins it might be a useful concept. It is now (since CHNOSZ version 0.9) the defaut mode for `diagram()` when working with proteins.

With the potential for calculating equilibrium activities of proteins comes the desire to compare these calculations to actual measurements!

4 Becoming Human

Let's look at some protein abundance levels in human blood plasma. First get going with the experimental data. In CHNOSZ is a table listing the upper limits of the intervals, or ranges, of protein abundances taken from figures available in Anderson and Anderson, 2002, 2003 [1, 2]. The protein abundances in the tables are in log10(pg/ml); let's convert that to molality. First locate the file with the abundance data. Then read it. Then identify the protein named "INS.C" and drop it from the list. The reason for doing so is that preliminary calculations show it is much more stable than any other protein in the list. It is therefore an interesting outlier in terms of relative stabilities of the proteins.

Then get the species indices of the proteins for thermodynamic calculations (with parameters based on amino acid compositions of the proteins listed in `thermo$protein` ... and doing so quietly, without console messages that would fill up a whole page here). Then calculate the masses of the proteins. Then convert log10(pg/ml) to log10(mol/L) (logarithm of molarity). The conversion from pg/ml to g/L involves a factor of 10^{-9} ($\frac{10^0\text{g}}{10^{12}\text{pg}} \times \frac{10^3\text{ml}}{10^0\text{L}}$); then to get molarity we divide by mass ($\frac{\text{g}}{\text{mol}}$).

```
> f <- system.file("extdata/abundance/AA03.csv", package = "CHNOSZ")
> pdata <- read.csv(f)
> pdrop <- which(pdata$name == "INS.C")
> pname <- pdata$name[-pdrop]
> iip <- info(paste(pname, "HUMAN", sep = "_"), quiet = TRUE)
```

```
protein: found HBA_HUMAN (C685H1071N187O194S3, 141 residues)
protein: found ALBU_HUMAN (C2936H4624N786O889S41, 585 residues)
protein: found IGHG1_HUMAN (C1612H2515N425O494S11, 330 residues)
protein: found TRFE_HUMAN (C3306H5134N912O1002S47, 679 residues)
protein: found FIBA_HUMAN (C3929H6065N1161O1312S25, 831 residues)
protein: found IGHA1_HUMAN (C1661H2604N446O518S17, 353 residues)
protein: found A2MG_HUMAN (C7193H11249N1901O2178S50, 1451 residues)
protein: found IGJ_HUMAN (C664H1071N187O227S9, 137 residues)
protein: found A1AT_HUMAN (C2001H3130N514O601S10, 394 residues)
protein: found CO3_HUMAN (C8234H13052N2224O2485S62, 1641 residues)
protein: found HPT_HUMAN (C1929H2984N520O586S16, 388 residues)
protein: found APOA1_HUMAN (C1241H1977N347O389S3, 243 residues)
protein: found APOB_HUMAN (C23073H36367N6077O6918S102, 4536 residues)
protein: found A1AG1_HUMAN (C966H1472N252O299S5, 183 residues)
protein: found APOA_HUMAN (C21525H32288N6144O6983S322, 4529 residues)
protein: found CFAH_HUMAN (C6003H9162N1638O1850S99, 1213 residues)
protein: found CERU_HUMAN (C5397H8101N1417O1627S38, 1046 residues)
protein: found CO4A_HUMAN (C3723H5880N1034O1143S24, 767 residues)
protein: found CFAB_HUMAN (C3664H5713N1019O1119S33, 739 residues)
protein: found TTHY_HUMAN (C617H950N160O193S2, 127 residues)
protein: found CO9_HUMAN (C2664H4154N734O841S33, 538 residues)
protein: found C1QA_HUMAN (C1045H1635N305O314S6, 223 residues)
protein: found CO8B_HUMAN (C2668H4122N758O815S37, 537 residues)
protein: found CO5_HUMAN (C5055H7960N1324O1494S41, 999 residues)
protein: found PLMN_HUMAN (C3848H5907N1099O1188S58, 791 residues)
protein: found IGHD_HUMAN (C1866H2926N524O571S13, 384 residues)
protein: found IC1_HUMAN (C2458H3933N641O757S18, 500 residues)
protein: found CO6_HUMAN (C4420H6891N1257O1406S71, 913 residues)
protein: found CO7_HUMAN (C3944H6095N1117O1244S64, 821 residues)
protein: found CFAI_HUMAN (C2773H4305N771O840S50, 565 residues)
protein: found RET4_HUMAN (C926H1416N260O285S10, 183 residues)
protein: found CO3.C3c_HUMAN (C1059H1701N287O309S6, 206 residues)
protein: found THBG_HUMAN (C1984H3113N507O589S19, 395 residues)
protein: found CO2_HUMAN (C3560H5583N1015O1073S41, 732 residues)
protein: found THRB_HUMAN (C2858H4405N805O890S32, 579 residues)
protein: found CRP_HUMAN (C1151H1746N282O332S6, 224 residues)
protein: found CFAB.Bb_HUMAN (C2548H4013N689O756S20, 505 residues)
protein: found CO3.C3a_HUMAN (C386H640N126O110S9, 77 residues)
protein: found FRIL_HUMAN (C885H1382N244O268S5, 174 residues)
protein: found CCL5_HUMAN (C342H526N94O97S5, 66 residues)
protein: found VTNC_HUMAN (C2304H3460N652O709S20, 459 residues)
protein: found MYG_HUMAN (C769H1215N209O221S4, 153 residues)
protein: found THYG_HUMAN (C13302H20602N3688O4090S159, 2749 residues)
protein: found TPA_HUMAN (C2569H3928N746O781S40, 527 residues)
protein: found CO5.C5a_HUMAN (C350H584N108O107S8, 74 residues)
```

```
protein: found ENOG_HUMAN (C2084H3308N568O650S13, 433 residues)
protein: found FETA_HUMAN (C2922H4614N790O899S40, 591 residues)
protein: found TNR1A_HUMAN (C1204H1883N333O374S28, 251 residues)
protein: found KLK3_HUMAN (C1162H1817N323O333S14, 237 residues)
protein: found PPAP_HUMAN (C1855H2835N477O542S16, 354 residues)
protein: found CEAM5_HUMAN (C3140H4871N867O1011S12, 651 residues)
protein: found MBP_HUMAN (C1404H2215N463O463S4, 304 residues)
protein: found TNNI1_HUMAN (C936H1576N284O277S10, 186 residues)
protein: found IL1RA_HUMAN (C754H1177N207O231S9, 152 residues)
protein: found CCL4_HUMAN (C338H510N86O107S5, 67 residues)
protein: found TNNT1_HUMAN (C1423H2304N428O449S7, 277 residues)
protein: found IL8_HUMAN (C397H648N114O111S4, 77 residues)
protein: found CCL3_HUMAN (C327H499N85O106S4, 66 residues)
protein: found TF_HUMAN (C1328H2068N342O411S6, 263 residues)
protein: found CSF3_HUMAN (C856H1360N226O249S8, 178 residues)
protein: found IFNA1_HUMAN (C851H1348N234O261S11, 166 residues)
protein: found IL2_HUMAN (C693H1120N178O203S7, 133 residues)
protein: found IL4_HUMAN (C653H1062N192O196S7, 129 residues)
protein: found TNFA_HUMAN (C778H1227N215O231S2, 157 residues)
protein: found IFNG_HUMAN (C723H1145N199O214S4, 138 residues)
protein: found IL1B_HUMAN (C773H1219N201O237S8, 153 residues)
protein: found IL12A_HUMAN (C989H1592N264O301S17, 197 residues)
protein: found IL10_HUMAN (C823H1302N228O244S11, 160 residues)
protein: found IL5_HUMAN (C588H958N160O174S3, 115 residues)
protein: found IL6_HUMAN (C909H1475N253O286S9, 183 residues)
```

```
> pmass <- element(thermo$obigt$formula[iip])$mass
> loga.expt <- logm <- log10(10^pdata$log10.pg.ml.[-pdrop]/10^9/pmass)
```

As implied by the "loga", we are assuming for the comparisons offered below that molarity (derived from the published abundance data) can be taken to be equal to molality and that molality can equated with chemical activity. The latter equality (the assumption of ideal behavior) especially should be subject to more scrutiny. We'll go ahead anyway and calculate, for ideality, the equilibrium activities of the proteins. First we need to calculate the total activity of residues from the experimental data, but to do that we need even more firstly the lengths of the proteins.

```
> pl <- protein.length(paste(pname, "HUMAN", sep = "_"))
> logares.tot <- sum(10^loga.expt * pl)
```

Our total activity (*not* the logarithm of it) of residues turns out to be about 200, which for our average protein length of 637 works out to about 0.3 for the average protein, if the total activity of residues could be attributed to that single average protein.

Now let's get down to the stuff CHNOSZ is made for. First define the basis species. Then define the species, being the proteins; for some reason (most likely the existence of factors in R) the "as.character" is needed to avoid an error. Then calculate the affinities of the formation reactions of the proteins. Then calculate the equilibrium activities, but don't plot them by themselves. Instead, use `revisit` to compare the equilibrium activities to the experimental abundances.

```
> basis("CHNOS+")
```

```
    C H N O S Z ispecies logact state
CO2 1 0 0 2 0 0       69     -3    aq
H2O 0 2 0 1 0 0        1      0   liq
NH3 0 3 1 0 0 0       68     -4    aq
H2S 0 2 0 0 1 0       70     -7    aq
O2  0 0 0 2 0 0     2691    -80   gas
H+  0 1 0 0 0 1        3     -7    aq
```

```
> species(as.character(pname), "HUMAN", quiet = TRUE)
> a <- affinity()
```

```
affinity: temperature is 25 C
energy.args: pressure is Psat
affinity: loading ionizable protein groups
subcrt: 93 species at 298.15 K and 1 bar (wet)
```

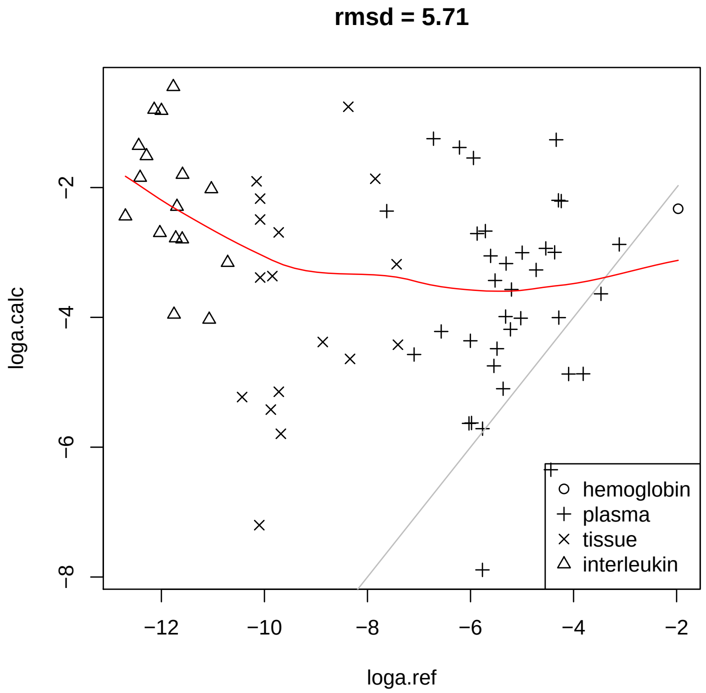
```
> d <- diagram(a, logact = logares.tot, do.plot = FALSE)
```

```
diagram: immobile component is protein backbone group
diagram: conservation coefficients are 141 585 330 679 831 353 1451 137 394 1641 388 243 4536 183 4529 1
diagram: using residue equivalents
diagram: log total activity of PBB (from argument) is 2.348706
```

```
> pch <- as.numeric(pdata$type)
> revisit(d, "rmsd", loga.ref = loga.expt, pch = pch)
```

```
revisit: calculating rmsd in 0 dimensions
```

```
> legend("bottomright", pch = unique(pch), legend = unique(pdata$type))
```

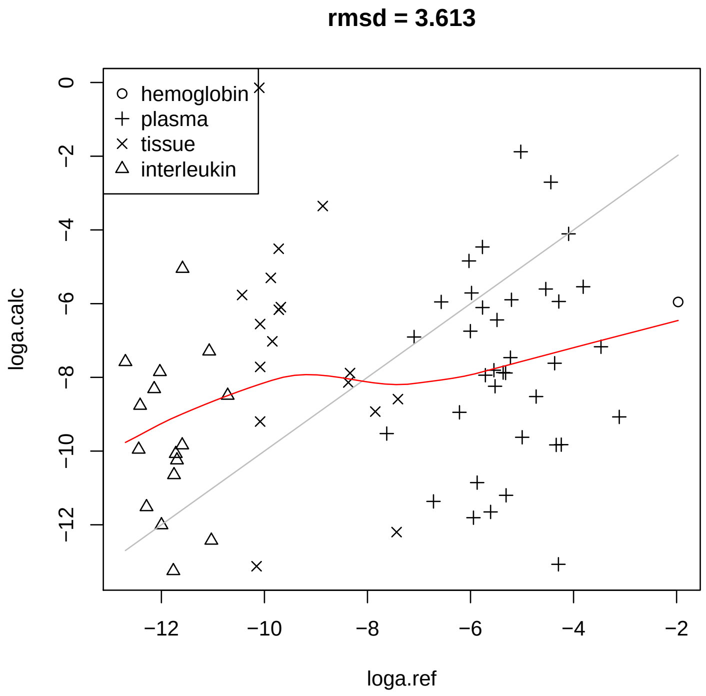


There seems to be almost no relation between the reference values and the calculated ones. But what if we increase the oxygen fugacity? $\log f_{O_{2(g)}} = -80$ might be appropriate for some subcellular conditions, or reduced hydrothermal systems. Blood is exposed to oxygen after all... let's try $\log f_{O_{2(g)}} = -60$.

```
> basis("O2", -60)
```

```
    C H N O S Z ispecies logact state
CO2 1 0 0 2 0 0       69     -3    aq
H2O 0 2 0 1 0 0        1      0   liq
NH3 0 3 1 0 0 0       68     -4    aq
H2S 0 2 0 0 1 0       70     -7    aq
O2  0 0 0 2 0 0     2691    -60   gas
H+  0 1 0 0 0 1        3     -7    aq
```

```
> a <- affinity()

affinity: temperature is 25 C
energy.args: pressure is Psat
affinity: loading ionizable protein groups
subcrt: 93 species at 298.15 K and 1 bar (wet)

> d <- diagram(a, logact = logares.tot, do.plot = FALSE)

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 141 585 330 679 831 353 1451 137 394 1641 388 243 4536 183 4529 1
diagram: using residue equivalents
diagram: log total activity of PBB (from argument) is 2.348706

> revisit(d, "rmsd", loga.ref = loga.expt, pch = pch)

revisit: calculating rmsd in 0 dimensions

> legend("topleft", pch = unique(pch), legend = unique(pdata$type))
```



Well it's still quite scattered. However, the RMSD has decreased considerably, the loess fit has a positive slope, and the dynamic ranges of the calculations and observations are more similar.

5 Comparison with expression profile in *E. coli*

Amino acid compositions of proteins in *Escherichia coli* are provided with CHNOSZ at `extdata/protein/ECO.csv.xz`. Protein abundances in the cytosol of *E. coli* reported by Ishihama et al., 2008 [6] are provided with CHNOSZ at `extdata/abundances/ISR+08.csv.xz`. We can use `get.expr()` to retrieve the abundance data for all or only selected proteins, and also add these proteins to CHNOSZ's inventory (`thermo$protein`) based on amino acid compositions from the `ECO.csv` file. First though we use `data(thermo)` to clear out the settings from the previous calculations.

```
> data(thermo)
> file <- system.file("extdata/abundance/ISR+08.csv", package = "CHNOSZ")
> expr <- get.expr(file, "ID", "emPAI", "ECO", list(description = "kinase"))
```

```
get.expr: searching for 36 entries... get.protein: KPY1 PPCK K6P1 KPY2 K6P2 were not matched
get.protein: found 31 of 36 proteins
add.protein: added 31 of 31 proteins
```

```
> range(expr$loga.ref)
```

```
[1] -8.029615 -2.597608
```

The result (`expr`) lists data for proteins where the `description` column of ISR+08.csv contains the term `kinase`. The list has elements named `id` (corresponding to the `ID` column of ISR+08.csv), `iprotein` (corresponding to the rownumber of the proteins in thermo$protein) and `loga.ref` (logarithm of activity, corresponding to the `emPAI` column of ISR+08.csv, scaled so that total activity of residues is unity). Note that the ID's of five of the 36 proteins that are described as "kinase" are not found in ECO.csv, so only 31 proteins are returned by the above call to `get.expr()`. The minimum and maximum values of the reference abundances are separated by over five orders of magnitude.

Now we can calculate the metastable equilibrium activities of the proteins, setting the total activity of residues to unity. We then use `revisit()` to make a plot and compute the root mean square deviation between the experimental and calculated relative abundances. Since the equilibrium activities of the proteins were only calculated at a single point, `revisit()` here makes a scatter plot. The colors reflect the average oxidation state of carbon of the proteins (red – more reduced, blue – more oxidized).

```
> basis("CHNOS+")
```

```
    C H N O S Z ispecies logact state
CO2 1 0 0 2 0 0       69     -3    aq
H2O 0 2 0 1 0 0        1      0   liq
NH3 0 3 1 0 0 0       68     -4    aq
H2S 0 2 0 0 1 0       70     -7    aq
O2  0 0 0 2 0 0     2691    -80   gas
H+  0 1 0 0 0 1        3     -7    aq
```

```
> a <- affinity(iprotein = expr$iprotein)
```

```
affinity: temperature is 25 C
energy.args: pressure is Psat
protein: found H2O_RESIDUE (H2O, 0 residues)
protein: found Ala_RESIDUE (C3H5NO, 1 residues)
protein: found Cys_RESIDUE (C3H5NOS, 1 residues)
protein: found Asp_RESIDUE (C4H5NO3, 1 residues)
protein: found Glu_RESIDUE (C5H7NO3, 1 residues)
protein: found Phe_RESIDUE (C9H9NO, 1 residues)
protein: found Gly_RESIDUE (C2H3NO, 1 residues)
protein: found His_RESIDUE (C6H7N3O, 1 residues)
protein: found Ile_RESIDUE (C6H11NO, 1 residues)
protein: found Lys_RESIDUE (C6H12N2O, 1 residues)
protein: found Leu_RESIDUE (C6H11NO, 1 residues)
protein: found Met_RESIDUE (C5H9NOS, 1 residues)
protein: found Asn_RESIDUE (C4H6N2O2, 1 residues)
protein: found Pro_RESIDUE (C5H7NO, 1 residues)
protein: found Gln_RESIDUE (C5H8N2O2, 1 residues)
protein: found Arg_RESIDUE (C6H12N4O, 1 residues)
protein: found Ser_RESIDUE (C3H5NO2, 1 residues)
```

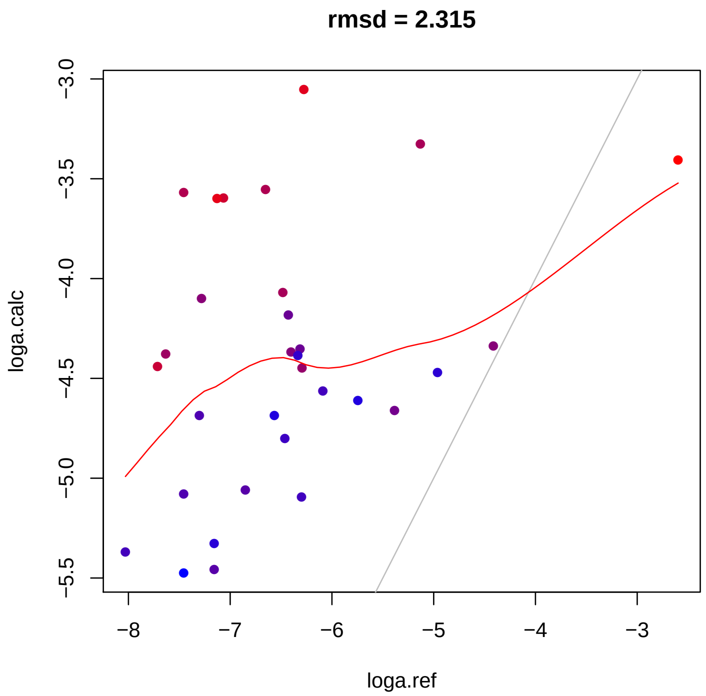
```
protein: found Thr_RESIDUE (C4H7NO2, 1 residues)
protein: found Val_RESIDUE (C5H9NO, 1 residues)
protein: found Trp_RESIDUE (C11H10N2O, 1 residues)
protein: found Tyr_RESIDUE (C9H9NO2, 1 residues)
affinity: loading ionizable protein groups
subcrt: 44 species at 298.15 K and 1 bar (wet)
```

```
> d <- diagram(a, logact = 0, do.plot = FALSE)
```

```
diagram: immobile component is protein backbone group
diagram: conservation coefficients are 387 400 820 502 214 315 420 566 173 143 347 241 382 227 367 207 3
diagram: using residue equivalents
diagram: log total activity of PBB (from argument) is 0
```

```
> tp <- thermo$protein[expr$iprotein, ]
> z <- ZC(protein.formula(tp))
> col <- rgb(max(z) - z, 0, z - min(z), max = diff(range(z)))
> revisit(d, "rmsd", loga.ref = expr$loga.ref, pch = 16, col = col)
```

```
revisit: calculating rmsd in 0 dimensions
```

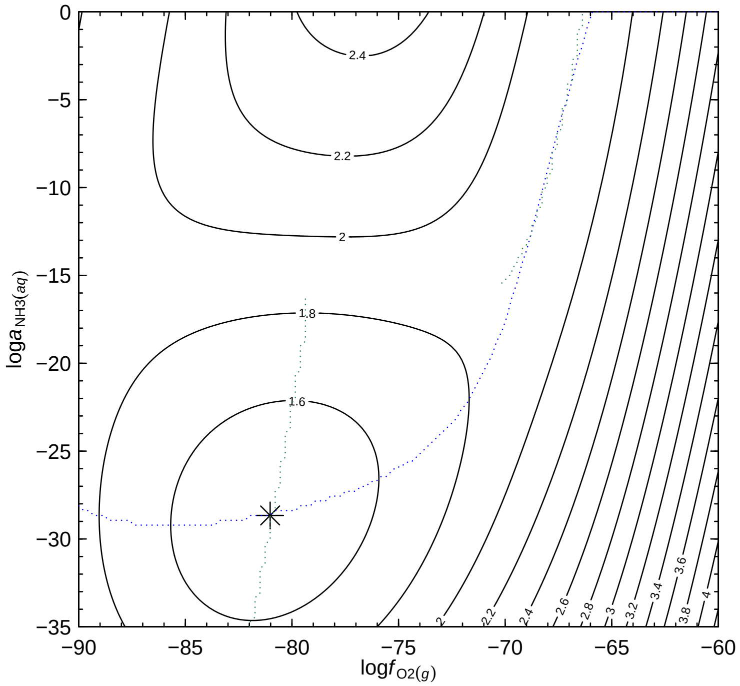


How can the correlation be improved? We can find where the RMSD minimizes as a function of a single variable. Or let's go for two variables ... note that we have to specify `mam=FALSE` in the call to `diagram()` in this case:

```
> a <- affinity(O2 = c(-90, -60), NH3 = c(-35, 0), iprotein = expr$iprotein)
```

```
affinity: temperature is 25 C
energy.args: pressure is Psat
energy.args: variable 1 is O2 at 128 increments from -90 to -60
energy.args: variable 2 is NH3 at 128 increments from -35 to 0
affinity: loading ionizable protein groups
subcrt: 44 species at 298.15 K and 1 bar (wet)
```

```
> d <- diagram(a, logact = 0, do.plot = FALSE, mam = FALSE)

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 387 400 820 502 214 315 420 566 173 143 347 241 382 227 367 207 3
diagram: using residue equivalents
diagram: log total activity of PBB (from argument) is 0

> r <- revisit(d, "rmsd", loga.ref = expr$loga.ref)

revisit: calculating rmsd in 2 dimensions
```



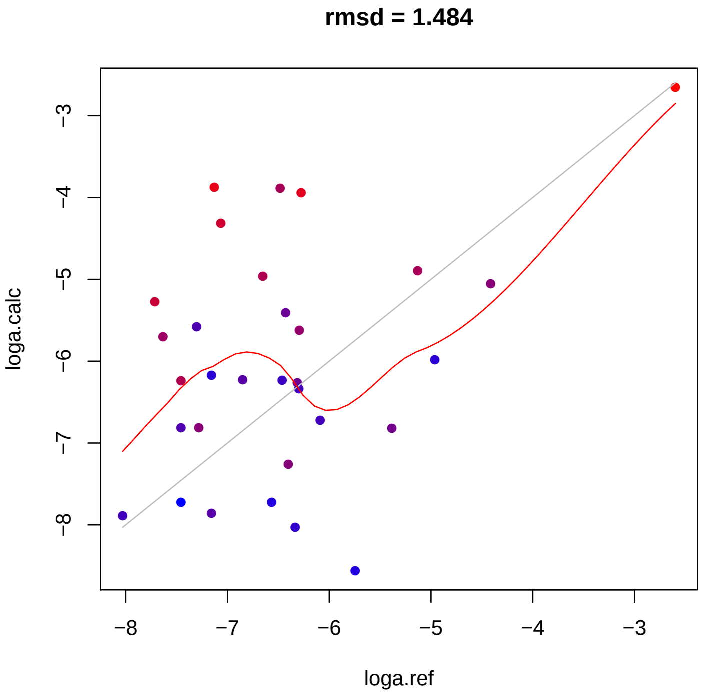
Now set the activities of the basis species where the minimum RMSD was found, calculate the affinities and equilibrium activities, and compare the results with the reference abundances.

```
> basis(c("O2", "NH3"), c(r$x, r$y))

    C H N O S Z ispecies    logact state
CO2 1 0 0 2 0 0       69  -3.00000    aq
H2O 0 2 0 1 0 0        1   0.00000   liq
NH3 0 3 1 0 0 0       68 -28.66142    aq
H2S 0 2 0 0 1 0       70  -7.00000    aq
O2  0 0 0 2 0 0     2691 -81.02362   gas
H+  0 1 0 0 0 1        3  -7.00000    aq

> a <- affinity(iprotein = expr$iprotein)

affinity: temperature is 25 C
energy.args: pressure is Psat
affinity: loading ionizable protein groups
subcrt: 44 species at 298.15 K and 1 bar (wet)

> d <- diagram(a, logact = 0, do.plot = FALSE)
```

```
diagram: immobile component is protein backbone group
diagram: conservation coefficients are 387 400 820 502 214 315 420 566 173 143 347 241 382 227 367 207 3
diagram: using residue equivalents
diagram: log total activity of PBB (from argument) is 0
```

```
> revisit(d, "rmsd", loga.ref = expr$loga.ref, pch = 16, col = col)
```

```
revisit: calculating rmsd in 0 dimensions
```



6 Summary

Using default settings, equilibrium activities of proteins are calculated in CHNOSZ by converting formation reactions of proteins to their per-residue equivalents, then using the Boltzmann distribution to transform the affinities of the formation reactions (in an equal-activity reference state) to equilibrium activities (an equal-affinity reference state).

The construction of 2-D predominance diagrams (for proteins or any other type of system) by default avoids calculating the equilibrium activities of species and instead identifies predominant species based on maximum affinity (after normalizing by the conservation coefficients). For systems of proteins, set `mam=FALSE` in `diagram()` to run the activity calculations if these values are needed, such as in the *E. coli* example above.

If oxygen fugacity is raised from its default nominal setting in CHNOSZ, the dynamic range of equilibrium activities calculated for proteins in human plasma becomes similar to the observed reference abundances of the proteins, and a slight positive correlation emerges. Equilibrium activities of kinases in *E. coli* cytosol have a dynamic range that is also similar to the observed abundances, but our findings so far imply going to a very low chemical potential of nitrogen (in terms of $\log a_{\mathrm{NH}_{3(aq)}}$) to minimize the overall deviation.

7 Document Information

Revision history

- 2009-11-29 Initial version (Calculating relative abundances of proteins)

- 2011-06-20 Add human and *E. coli* comparisons

R session information

```
> sessionInfo()

R version 2.13.1 (2011-07-08)
Platform: x86_64-slackware-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US       LC_NUMERIC=C         LC_TIME=en_US
 [4] LC_COLLATE=C         LC_MONETARY=C        LC_MESSAGES=en_US
 [7] LC_PAPER=en_US       LC_NAME=C            LC_ADDRESS=C
[10] LC_TELEPHONE=C       LC_MEASUREMENT=en_US LC_IDENTIFICATION=C

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
[1] CHNOSZ_0.9-6
```

References

[1] N. L. Anderson and N. G. Anderson. The human plasma proteome - history, character, and diagnostic prospects. *Molecular & Cellular Proteomics*, 1(11):845–867, November 2002. doi: 10.1074/mcp.R200007-MCP200.

[2] N. L. Anderson and N. G. Anderson. The human plasma proteome: History, character, and diagnostic prospects (vol 1, pg 845, 2002). *Molecular & Cellular Proteomics*, 2(1):50–50, January 2003. doi: 10.1074/mcp.A300001-MCP200.

[3] J. M. Dick. Calculation of the relative metastabilities of proteins using the CHNOSZ software package. *Geochem. Trans.*, 9:10, 2008. doi: 10.1186/1467-4866-9-10.

[4] J. M. Dick and E. L. Shock. Calculation of the relative chemical stabilities of proteins as a function of temperature and redox chemistry in a hot spring. *PLoS ONE*, 6(8):e22782, 2011. doi: 10.1371/journal.pone.0022782.

[5] J. M. Dick, D. E. LaRowe, and H. C. Helgeson. Temperature, pressure, and electrochemical constraints on protein speciation: Group additivity calculation of the standard molal thermodynamic properties of ionized unfolded proteins. *Biogeosciences*, 3(3):311 – 336, 2006. doi: 10.5194/bg-3-311-2006.

[6] Y. Ishihama, T. Schmidt, J. Rappsilber, M. Mann, F. U. Hartl, M. J. Kerner, and D. Frishman. Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics*, 9:102, FEB 27 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-102.

[7] J. S. Seewald. Mineral redox buffers and the stability of organic compounds under hydrothermal conditions. *Mat. Res. Soc. Symp. Proc.*, 432:317 – 331, 1996. doi: 10.1557/PROC-432-317.