Calculating relative abundances of proteins

Jeffrey M. Dick

November 30, 2009

Abstract

In previous versions of CHNOSZ, the default argument `residue=FALSE` in the `diagram` function set the code to use an equation-solving approach to calculating the equilibrium relative abundances of chemical species, which can be inorganic molecules or proteins. For inorganic chemical species that are not large polymers, it make sense to use this setting. However, it is relatively slow, and gives activities of proteins that are not compatible with their coexistence in a metastable equilibrium state. By setting `residue=TRUE`, the function effectively rewrites the formation reactions for the residue equivalents of the proteins (or any other species), which leads to equilibrium activities that are compatible with the coexistence of proteins in a metastable equilibrium state. Also, because the number of moles of the conservant in the reactions is always unity, it permits a much less computationally intense approach for calculating the relative abundances of proteins that is based on the Maxwell-Boltzmann distribution. Therefore, in CHNOSZ beginning with version 0.9, the default setting of the `residue` argument is changed to `TRUE` when the system is composed entirely of proteins. The resulting changes in the species distributions on 1-D speciation diagrams as well as on 2-D predominance diagrams are described.

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 second purpose of the function, and is the subject of this document.

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 equates to molality. Since molality is a measure of concentration, calculating the chemical activities, or the molalities, is just a way of calculating the relative abundances of species.

2 Calculations at a single point

Here we discuss two procedures for calculating relative abundances of species. The first, the equation-solving approach, is described in Dick, 2008[1] and is summarized below using the same example given in the paper. The second, the Maxwell-Boltzmann distribution, is also described by example below. We show (by example) that the two approaches are equivalent when the formation reactions of residue equivalents of proteins are used.

[1] Dick, J. M. Calculation of the relative metastabilities of proteins using the CHNOSZ software package. *Geochemical Transactions*, 9, 10, 2008. `http://dx.doi.org/10.1186/1467-4866-9-10`

2.1 Equation-solving approach

The next two sections give examples of calculating the relative abundances of proteins using an equation-solving approach. The Appendix explains the procedure actually used in `diagram` to solve equations of this type. 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 power.

2.1.1 Whole proteins

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

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

and

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

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)

CHNOSZ version 0.9 (2009-11-30)
thermo: loaded 1997 aqueous, 3089 total species to thermo$obigt
thermo: loaded 5264 proteins to thermo$ECO
thermo: loaded 6717 proteins to thermo$SGD
thermo: loaded 4155 localizations and 3570 abundances to thermo$yeastgfp

> 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     2852    -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)

> species()

   CO2  H2O NH3 H2S      O2 H+ ispecies logact state      name
1 2575 1070 645  11 -2668.0  0     3090     -3    aq CSG_METVO
2 2555 1042 640  14 -2643.5  0     3091     -3    aq CSG_METJA
```

Although the basis species are defined, the temperature is not yet specified, so it is not immediately possible to assign charges to the proteins. That is why the coefficient on H^+ is zero in the output above. To see what the 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: 3090 refers to CSG_METVO, C2575H4097N645O884S11 aq (BBA+03).
info: 3091 refers to CSG_METJA, C2555H4032N640O865S14 aq (BBA+03).
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 to the group additivity and equations-of-state parameters from Dick et al., 2006[2]. 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

$`3090`
[1] 107.6774

$`3091`
[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[3],

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

for Reactions 3 and 4 as

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

and

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

[2]Dick, J. M., LaRowe, D. E. and Helgeson, H. C. Temperature, pressure and electrochemical constraints on protein speciation: Group additivity calculation of the standard molal thermodynamic properties of ionized unfolded proteins. *Biogeosciences*, 3, 311–336, 2006. `http://www.biogeosciences.net/3/311/2006/bg-3-311-2006.html`

[3]2.303 is used here to stand for $\ln 10$.

Next we must choose the conservant, or immobile component, of the system. For proteins, it makes sense 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:

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

It also follows from our choice of immobile component that at equilibrium, $\boldsymbol{A_3}/553 = \boldsymbol{A}_4/530$. Let us now define $\boldsymbol{A}_3^* \equiv \boldsymbol{A}_3 + 2.303RT \log a_{\text{CSG\_METVO}} = 2.303RT \times 104.6774$ and $\boldsymbol{A}_4^* \equiv \boldsymbol{A}_4 + 2.303RT \log a_{\text{CSG\_METJA}} = 2.303RT \times 314.1877$ to write, for equilibrium,

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

and

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

where $\boldsymbol{A} = \boldsymbol{A}_3/553 = \boldsymbol{A}_4/530$; the values in the denominators define the conservation coefficients. Now we have three equations with three unknowns. The solution can be displayed in CHNOSZ as follows. The argument `residue=FALSE` overrides the now-default setting for `diagram` when proteins are the species of interest and instructs it to use the function named `abundance.old`, which implements the equation-solving strategy described in the next section.

```
> 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

$`3090`
[1] -225.9512

$`3091`
[1] -2.689647
```

Combining these values with Eqs. (16) or (17) gives us $\boldsymbol{A}_3/2.303RT = 0.5978817$.

As noted in the paper, activities that vary by such extreme magnitudes make it seem like proteins are very unlikely to coexist in metastable equilibrium. Later we explain the concept of using residue equivalents of the proteins to deal with this strange result.

2.1.2 Implementing the equation-solving 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: $\boldsymbol{A} = \left(\boldsymbol{A}_i^* - 2.303RT \log a_i\right)/n_{\text{ic},i}$, where $n_{\text{ic},i}$ 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, $n_{\text{ic},i}$ 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

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

and

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

Note that 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 what these formation reactions look like:

```
> 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

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

Note that 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 substitute $\boldsymbol{A}$ 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 $\boldsymbol{A}^*_{13} = \boldsymbol{A}_{13} + 2.303RT \log a_{\mathrm{CSG\_METJA(residue)}}$. Since we are following the example given in the paper, which uses 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 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 $a_{\text{CSG\_METVO(residue)}} = 0.3065982$, $a_{\text{CSG\_METJA(residue)}} = 0.7764018$ 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 hypothesis that some proteins coexist in metastable equilibrium. However, the approach just described is not used in CHNOSZ when `residue=TRUE`. Instead, the Maxwell-Boltzmann distribution, described next, it used, because it is much faster and leads to the same results.

2.2 Maxwell-Boltzmann distribution

It turns out that the equation-solving approach is slow. This is a major hindrance for any project that aims to predict the relative abundances of proteins in ecosystems as a function of multiple geochemical variables such as temperature and oxidation-reduction potential. Fortunately, there is a speedy way to get the relative abundances of proteins when the formation reactions are written for residue equivalents and residues are conserved. It comes from the Maxwell-Boltzmann distribution, which can be written in our notation as

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

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 $a_{13} = 0.7764018$.

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 `abundance.new`.

```
> 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 abundances of any number of proteins.

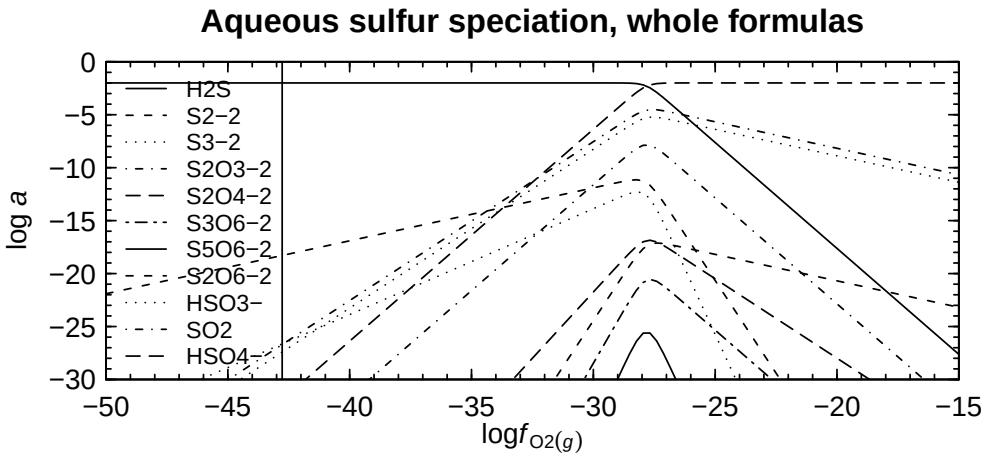
3 Calculations as a function of a single variable

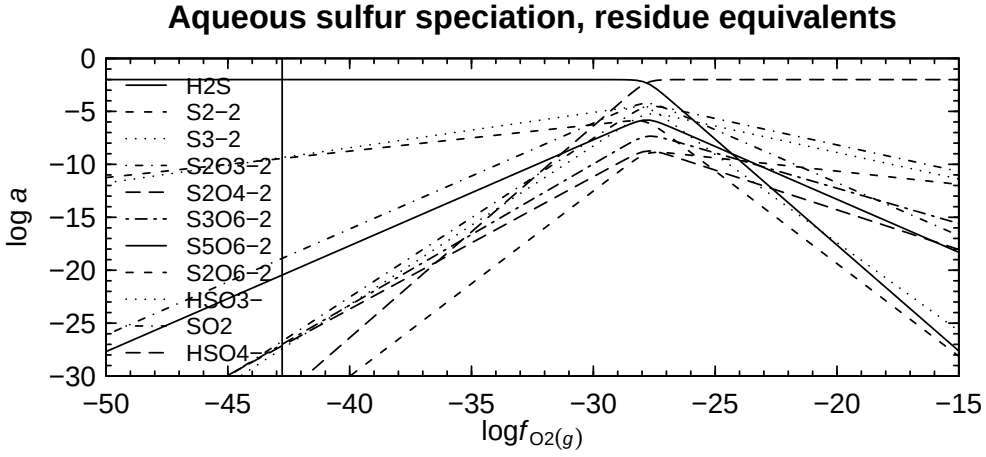
Unlike the application of the Maxwell-Boltzmann distribution described above, the equation-solving approach can be generalized for conservation coefficients that contain values other than unity, so it is appropriate for many mineral and inorganic systems where the immobile component has different coefficients in the formulas of the species of interest. For example, consider a system like that described by Seewald, 1997[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     2852    -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     2852    -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-"))
> species()

   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     1043     -3    aq S2O4-2
6    0  -2   0   3 4.0 -2     1048     -3    aq S3O6-2
7    0  -4   0   5 5.0 -2     1050     -3    aq S5O6-2
8    0  -1   0   2 3.5 -2     1047     -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)
```

[4]Seewald, J. S., 1997. Mineral redox buffers and the stability of organic compounds under hydrothermal conditions. *Mat. Res. Soc. Symp. Proc.*, 432, 317-331, 1997. http://lucy.mrs.org/meetings/spring96/Program/S.S96.html

```
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")
```

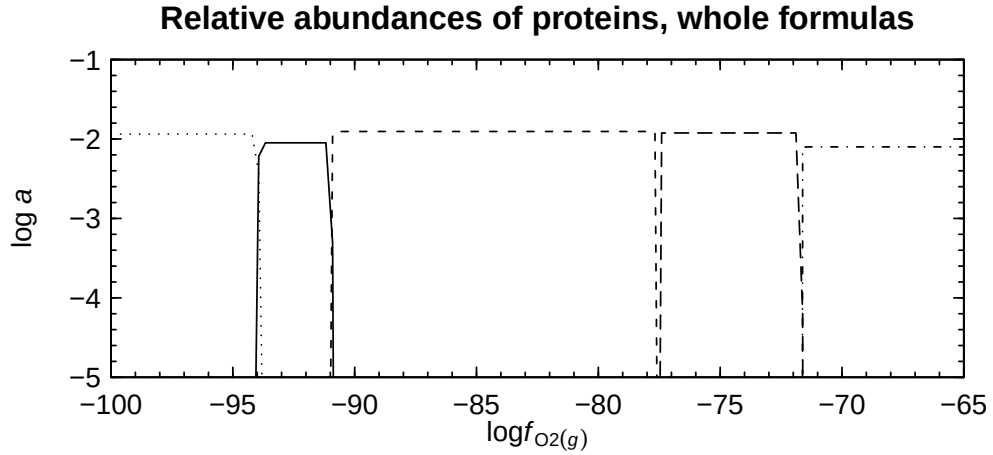


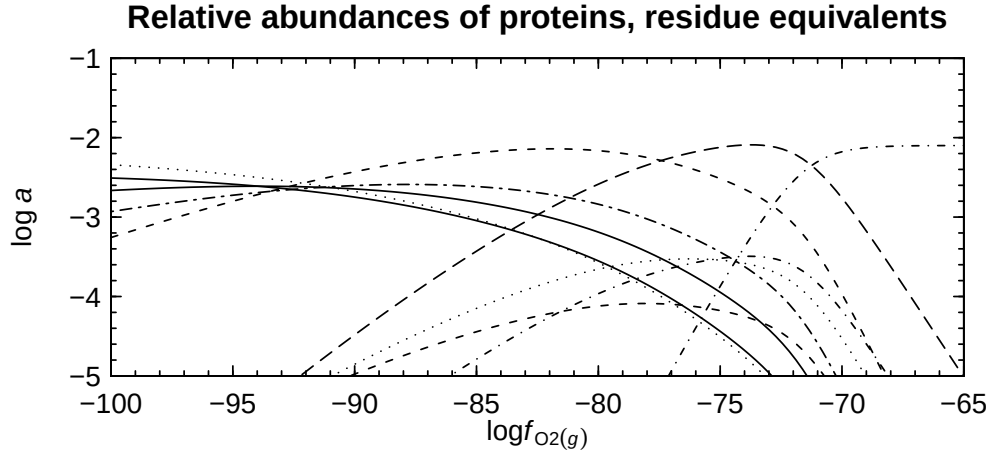


The first diagram is very similar to the one shown by Seewald, 1997, but the second (where we have set `residue=TRUE`) is quite different. There, the function was told to rewrite the formation reactions of the aqueous sulfur species for their residue equivalents (where the number of "residues" in each species is the coefficient of the immobile component, in this case H_2S, in the formation reaction) in the same way the formation reactions for the proteins were rewritten above.

An analogous comparison for proteins was given in Dick, 2008 and is reproduced below (though, without labels on the figures; a finer rendition of the `residue=TRUE` diagram can be found in the examples in the CHNOSZ documentation).

```
> 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     2852    -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)

> 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 = "Relative abundances 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 = "Relative abundances of proteins, residue equivalents")
```





4 Calculations as a function of two variables

How does `diagram` make a predominance diagram? Usually it uses the maximum affinity method, which is just a recognition of Eq. (19): the species with the highest affinity of formation for a reference (non-equilibrium) chemical activity is the one with the highest activity at equilibrium. The maximum affinity method coded in `diagram` is sensitive to the `residue` and `as.residue` arguments. This is shown below by the first three calls to `diagram` below, which use the default setting of `mam=TRUE`. Only the last one uses `mam=FALSE`, where the relative abundances of all four proteins are calculated using the Maxwell-Boltzmann distribution.

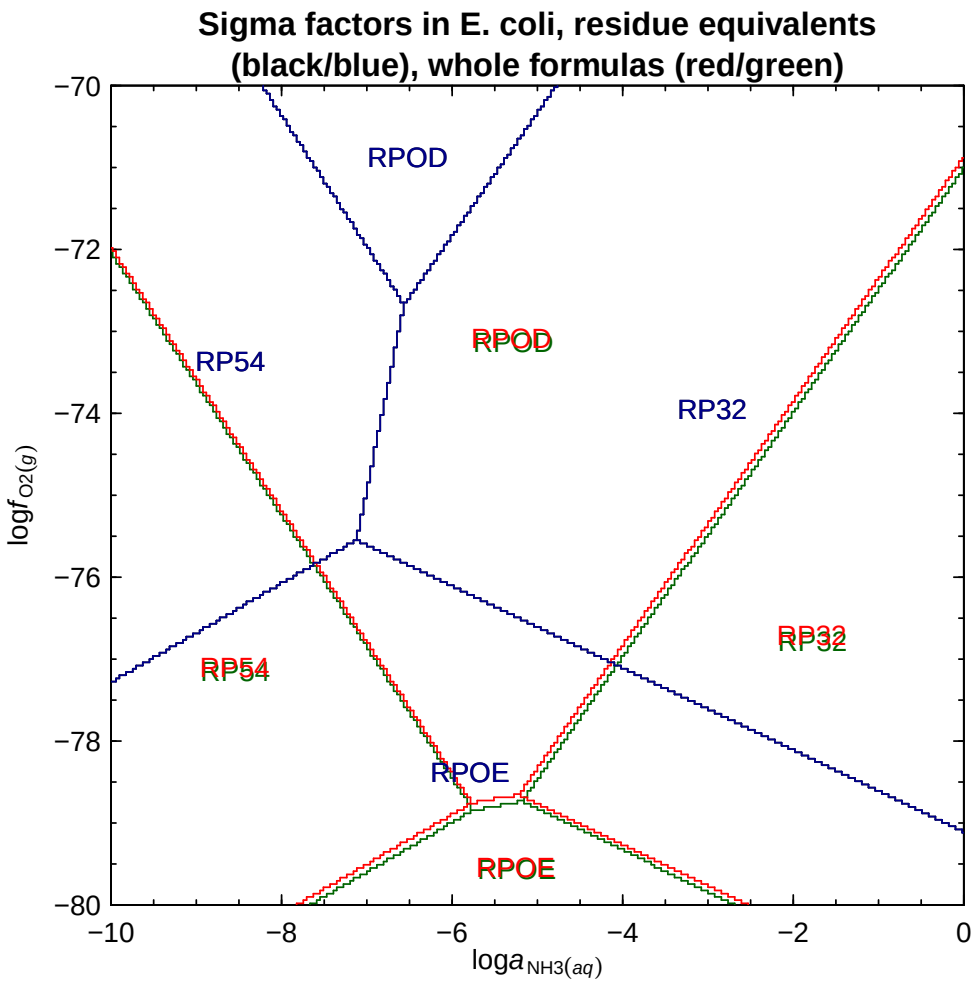
```
> basis("CHNOS")
```

```
    C H N O S ispecies logact state
CO2 1 0 0 2 0       69     -3    aq
H2O 0 2 0 1 0        1      0   liq
NH3 0 3 1 0 0       68     -4    aq
H2S 0 2 0 0 1       70     -7    aq
O2  0 0 0 2 0     2852    -80   gas
```

```
> species(c("RPOE", "RP32", "RP54", "RPOD"), "ECOLI")
```

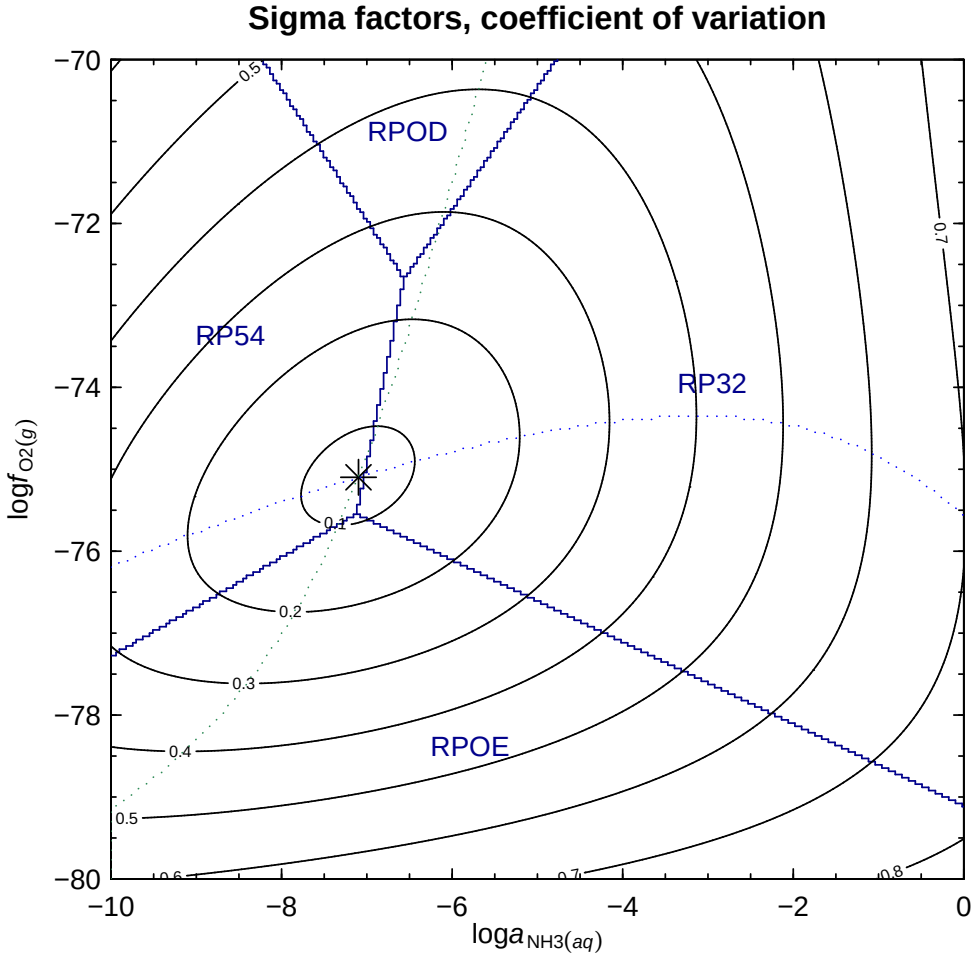
```
protein: found RPOE_ECOLI (C960H1541N271O291S5, 191 residues)
protein: found RP32_ECOLI (C1423H2264N414O434S11, 284 residues)
```

```
protein: found RP54_ECOLI (C2365H3791N649O766S13, 477 residues)
protein: found RPOD_ECOLI (C3042H4912N856O993S28, 613 residues)

> a <- affinity(NH3 = c(-10, 0, 256), O2 = c(-80, -70, 256))

affinity: temperature is 25 C
energy.args: pressure is Psat
energy.args: variable 1 is NH3 at 256 increments from -10 to 0
energy.args: variable 2 is O2 at 256 increments from -80 to -70
subcrt: 9 species at 298.15 K and 1 bar (wet)

> diagram(a, color = NULL)

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 191 284 477 613
diagram: using residue equivalents

> diagram(a, as.residue = TRUE, add = TRUE, col = "darkgreen")

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 191 284 477 613
diagram: using residue equivalents

> diagram(a, residue = FALSE, add = TRUE, col = "red")

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 191 284 477 613

> d <- diagram(a, mam = FALSE, add = TRUE, col = "darkblue")

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 191 284 477 613
diagram: using residue equivalents
diagram: log total activity of PBB (from species) is 0.1945143

> t1 <- "Sigma factors in E. coli, residue equivalents"
> t2 <- "(black/blue), whole formulas (red/green)"
> title(main = paste(t1, t2, sep = "\n"))
```



The last call to `diagram` gives the same result as the first one. So, the setting of `mam` does not matter for making predominance diagrams. Its default is `FALSE` because the calculation of relative abundances either takes a little bit longer (for `residue=TRUE`) or a lot longer (for `residue=FALSE`). But when we set `mam=TRUE` for the 2-dimensional predominance diagram, we can take the calculated relative abundances and make other plots, such as one showing the coefficient of variation of the activities of the proteins. This can be useful for a deeper understanding of the relative stabilities of the proteins, as it shows the conditions where their coexistence is most probable, from a thermodynamic standpoint.

```
> diagram(a, col = "darkblue", color = NULL)

diagram: immobile component is protein backbone group
diagram: conservation coefficients are 191 284 477 613
diagram: using residue equivalents

> draw.diversity(d, add = TRUE)
> title(main = "Sigma factors, coefficient of variation")
```



5 Summary and Recommendations

The equation-solving approach is used when `residue=FALSE`, which is the default setting in `diagram` for systems other than proteins. This makes sense for many inorganic and mineral systems, since it permits conservation vectors that have values other than unity. When studying systems of proteins (or possibly other biomacromolecules), set `residue=TRUE` to apply the Maxwell-Boltzmann distribution to formation reactions written for residue equivalents. This is now the default setting in `diagram` for systems composed entirely of proteins. One drawback of the latter is that it appears to be limited to cases where the conservation coefficients are all unity. However, it is computationally much less expensive than the equation-solving approach.

The construction of 2-D predominance diagrams (for proteins or any other type of system) by default avoids calculating the relative abundances of species and instead uses the maximum affinity method, indicated by `mam=TRUE`. For systems of proteins, set `mam=FALSE` to run the abundance calculations if these values are desired.