Revised: 8/9/00

Interpretation of Molecular Descriptors in QMF Files.

(Prof. Curt M. Breneman, RPI)

Analysis and Discussion of the Descriptor Fields

In the spirit of efficiency, the descriptors common to all kinds of QMF file formats will be discussed first. It should be noted that the **Volume** descriptor is only found in the Molecular section of the file formats. Its meaning is the same as similar descriptors computed using most modern molecular modeling programs. Molecular **Volume** is most often associated with hydrophobic effects, and tends to be correlated to the energy required to "dig a hole" in the solvent medium for the molecule. This is the sum of energies required to break existing non-covalent interactions between solvent molecules, as well as the desolvation energies of the binding site with which the molecule might interact. In the case of solution binding and molecular recognition, the desolvation energy of the solute molecule is also related to its **Volume**.

Molecular, regional or atomic **Surface Area** can be used in place of **Volume** descriptors in many cases, since the two can be highly correlated for analogous sets of molecules. For example, if the molecular shapes are similar through out a given dataset, these two descriptors are likely to be highly correlated.

The next type of index which is included in all current QMF file formats are derivatives of **SIDel(Rho)N**. This is an example of a vector property which has been scalarized by taking its dot product with respect to each local surface normal vector on the polygonalized molecular electron density isosurface (0.002 electrons / Bohr^3). All of the surface-derived indices are taken with respect to this locus of points surrounding each molecule. This surface is also used for defining the **Volume** and **Surface Area** properties mentioned in the previous section. As with the many other surface properties which are described in this document, several kinds of descriptors are provided which yield alternate ways of representing the surface distribution of the property. In this case, the basic property being described, **Del(Rho)N**, is Surface Integrated (origin of the "SI" prefix) over the region being examined -- This is either an atomic surface, a regional surface (usually a sum of atomic surfaces, but not necessarily), or a molecular surface. The integration occurs by defining the molecular surface as a set of triangular polygons, from which a set of two in-plane vectors can be defined. These vectors are then crossed, and a surface normal vector determined. The cross-product procedure also gives a measure of the surface area of the triangle. Next, the vector property to be examined is computed at each of the three triangle vertices. In the case of **SIDel(Rho)N**, the gradient vector of the electron density is computed at each point. These three gradient vectors are then dotted with the computed surface normal vector, and the three scalar values averaged. This scalar value is then multiplied by the surface area of the triangular surface element, and the procedure is repeated for all surface polygons on the structural region being considered. A related set of values are found in the **Del(Rho)NMin,**

Del(Rho)NMax and **Del(Rho)NIA** descriptors. The first two of these are simply extrema of the **Del(Rho)N** scalar values over the Atom, Region or Molecule under consideration. **Del(Rho)NIA** is the integral average of the scalarized electron density gradient normal to molecular surface. It can be considered a "surface area normalized" version of **SIDel(Rho)N**. For molecules of similar size, these two descriptors are often correlated. Less correlation between these indices are seen for molecules with different kinds of surface-exposed heteroatoms and are of different sizes and shapes. The next ten values represent histogram areas. They are: **Del(Rho)NA1, Del(Rho)NA2, Del(Rho)NA3, Del(Rho)NA4, Del(Rho)NA5,Del(Rho)NA6, Del(Rho)NA7, Del(Rho)NA8, Del(Rho)NA9,** and **Del(Rho)NA10**. These histogram bins are counted in units of square atomic units (Bohr), and represent the surface area occupied by each range of the property-weighted surface polygons over an Atom, Region or Molecule. For all Atomic, Regional or Molecular histograms, the starting value and "bin width" of these regions were all derived on the basis of a large number of observations, and the end bins are set to incorporate all of the surface polygons with property ranges outside the normal bin scale. In the case **Del(Rho)N**, the values cannot be positive, since the electron density gradient is always negative when dotted into the surface normal leading away from the molecule. The bin range starts at -$4x10^{-4}$ and ends at 0.0 with bin widths of 0.00005. The **Del(Rho)N** descriptor fields have been implicated in distinguishing "soft" regions of polarizable electron density from more tightly held regions. For example, the values of **Del(Rho)N** are much smaller over electron-rich pi systems and aromatic rings than over polarized or electron deficient alkyl carbons. Large negative values of **Del(Rho)N** functions or lower numbered bin populations (such as A1-A3) indicate that the electron density of the underlying molecular regions are more tightly held and less likely to extend very far from the molecule.

The next set of fields common to all QMF formats is that of the **SIDel(K)N, Del(K)Min, Del(K)Max, Del(K)IA, Del(K)NA1, Del(K)NA2, Del(K)NA3, Del(K)NA4, Del(K)NA5, Del(K)NA6, Del(K)NA7, Del(K)NA8, Del(K)NA9** and **Del(K)NA10** descriptors. As in the previous section, the prefix on the **SIDel(K)N** descriptor indicates that this index is a "surface integral" of the rate of change of the **K** electronic kinetic energy density normal to and away from the molecular surface. The **K** energy density is inherently negative, and so is the **Del(K)N** scalar function range. Consequently, the Atomic, Regional and Molecular histogram bin range begins at $-1.5x10^{-4}$ and ends at $2.5x10^{-4}$ with a bin range size of $5x10^{-5}$. The extrema and integral averages of this function are computed as described earlier, and are available for Atoms, Regions and Molecules. The **Del(K)N** family of descriptors have shown dominance in describing differences in the polarizability and hydrophobicity of molecular regions. More negative ranges of this function (or high populations of the lower bins A1-A3) are believed to indicate that the Atom, Region or Molecule is more hydrophobic and also less susceptible to electrophilic attack. Graphical observation of the distribution of this function on a number of molecular surfaces suggest that this is a very rapidly changing function over a small range of surface area, so this might give rise to numerical errors in its determination. More data is required to firm up this analysis.

The next set of fields to be discussed are related to the previous one, and include the **SIK, SIKMin, SIKMax, SIKIA, SIKA1, SIKA2, SIKA3, SIKA4, SIKA5, SIKA6, SIKA7, SIKA8, SIKA9,** and **SIKA10** descriptors. As before, the **SIK, SIKMin, SIKMax,and SIKIA** correspond to the full surface integral of the **K** kinetic energy density, as well as the extrema and integral average of this distribution. This is a rather smooth function over the surface of a typical molecule, and is sometimes complementary to the **G** electronic kinetic energy density distribution. Since the **K** function is always negative, the surface integral and integral average are always negative. The histogram range of this function is from -2.0x10^{-4} to 0.0 with a bin size of 2.5x10^{-5}. The **K** energy density is most negative in those portions of space where there is a "local concentration" of negative charge. This also corresponds to areas of negative Laplacian values, since imbalances of K and G electronic kinetic energy densities are responsible for non-zero Laplacian values. Such "negative Laplacian peaks" are usually seen within 0.25 - 0.4 Angstroms from an electron donor atom -- much inside the molecular Van der Waals surface chosen for this analysis. Nevertheless, "shadows" of these internal extrema are often present on the molecular surface and are therefore subject to analysis. These surface manifestations of the internal Laplacian peaks are often of the opposite sign as that of the actual peak as a result of Laplacian normalization. Consequently, slightly more positive (less negative) regions of surface values of **K** often indicate the presence of Bronsted bases.

Following the **K**-derived descriptors, the G-derived ones are present as: **SIDel(G)N, Del(G)NMin, Del(G)NMax, Del(G)NIA, Del(G)NA1, Del(G)NA2, Del(G)NA3, Del(G)NA4, Del(G)NA5, Del(G)NA6, Del(G)NA7, Del(G)NA8, Del(G)NA9,** and **Del(G)NA10**. As in the **K** case, the **SIDel(G)N, Del(G)NMin, Del(G)NMax,** and **Del(G)NIA** descriptors represent the surface integral and integral average of the scalarized gradient of the G density normal to and away from the molecular surface, as well as the extrema of the gradient functions themselves. The histogram bins for this surface distribution begins at –8 x10^{-4} and go up to 0.0 in 1x10^{-4} increments. The **G** kinetic energy density is always positive, which makes this gradient function always negative as the **G** density falls off when one moves away from the molecular surface. The **Del(G)N** family of descriptors have been seen together with **Del(Rho)N** functions in correlation models of dispersion interactions. These two families of functions are likely to show up in good correlation models of weak non-bonded interactions.

In a manner corresponding with the **K** functions, the **SIG, SIGMin, SIGMax, SIGIA, SIGA1, SIGA2, SIGA3, SIGA4, SIGA5, SIGA6, SIGA7, SIGA8, SIGA9,** and **SIGA10** functions are derived from the **G** electronic kinetic energy density as calculated on the molecular surface. In each case, the surface integral is computed in the manner described in the **Del(Rho)N** section, and then analyzed to find the extrema and the integral averages of the Atom, Region or Molecule of interest. The histogram bin values begin at 5 x10^{-6} and proceed through 7.25 x 10^{-4} in steps of 9 x 10^{-5}. Note: As in all of the Atomic, Regional and Molecular histograms, the sum of the bin populations is equal to the total surface area of the Atom, Region or Molecule. As with the associated **G** energy densities, the **SIG**-derived descriptors are usually associated with differences in donor/acceptor activities. There is also some relationship between **G** densities and hydrophobicities.

The next set of descriptors are of a more "traditional" nature, in that they are constructed to describe the electrostatic potential distribution on a molecular surface in a new way. The **SIEP, SIEPMin, SIEPMax, SIEPIA, SIEPA1, SIEPA2, SIEPA3, SIEPA4, SIEPA5, SIEPA6, SIEPA7, SIEPA8, SIEPA9,** and **SIEPA10** descriptors carry information about the electrostatic potential on the surface of an Atom, Region or Molecule. As in all of the previous surface integral-related descriptors, the **SIEP, SIEPMin, SIEPMax,** and **SIEPIA** indices refer to the overall surface integral of electrostatic potential as well as the (surface area normalized) integral average of the same quantity in addition to the extrema of the property-weighted surface integral values. The histogram bins range from $-8x10^{-2}$ to $8x10^{-2}$ in $2x10^{-2}$ steps. Electrostatic potential has been implicated in many molecular properties, including acid-base interactions, solvation behavior and pKa correlations. Many other useful correlations into the behavior of reactive molecules and unusual solvents have been analyzed using features of the molecular electrostatic potential distribution.

The electrostatic subset of the GIPF parameters described by Prof. Peter Politzer are also included in all versions of the QMF output. These descriptors, **piV, sigmaPV, sigmaNV, sumsigma,** and **sigmanew** describe several aspects of the surface electrostatic potential distribution. They are strictly Molecular descriptors, and do not have analogous Atomic or Regional counterparts. The **piV** descriptor, actually known in the Politzer literature as Π_v, is a polarity parameter describing the extent of polarity of a molecule. The **sigmaPV** and **sigmaNV** parameters are the standard deviations of the positive and negative portions of the surface electrostatic potential distribution, respectively. The **sigmanew** parameter is yet another electrostatic potential distribution width descriptor. These parameters, in conjunction with a volume-normalized polarizabilitiy parameter and a local average ionization potential minimum have been used to successfully model a large number of molecular properties. It is interesting to note that these parameters do not often show up in the top models of PLS regression MOMs in any of the datasets run to date. Their latent data appears to have been overshadowed by other indices which correlate more directly with the non-bonded intermolecular interactions that have been studied by our group so far.

An example of a more direct form of electrostatic potential descriptor can be found in the **EP1, EP2, EP3, EP4, EP5, EP6, EP7, EP8, EP9,** and **EP10** histogram bins. These bins are not found in the Atomic or Regional the QMF formats, but are present in the Molecular format sections of both “Old” and “New” QMF files. These bins directly represent the scalar electrostatic potential values on the surface of the molecules, and begin at $-4x10^{-3}$ atomic units and range up to $4x10^{-3}$ atomic units in increments of $1x10^{-3}$ atomic units. As in the **SIEP** form of this function, surface areas in each of the bins provide information about the distribution of the electrostatic potential over the molecular surface. These descriptors are often found in the best models of hydrogen bonding systems and in regressions concerning polar or dipolar molecules. Donor / Acceptor behavior is also modeled well using these histogram descriptors. The drawbacks of this kind of descriptor over those in the GIPF system are that more descriptors are needed in a given model to represent a subtle change in a distribution. On the other hand, the GIPF electrostatic potential descriptors give only a crude description of the shape of the surface potential. This is one argument for using both kinds of descriptors together in systems where electrostatic effects are likely to be important.

One of the most interesting and under-exploited of the GIPF parameters is the Local Average Ionization Potential, called “I-bar” in the Politzer literature but is called the “Politzer Ionization Potential” or “PIP” in our group. In keeping with the other classes of surface descriptors, there are two main types of PIP indices: The first type indicates extrema and average value descriptors **PIPMin, PIPMax,** and **PIPAvg,** and the second type is an extended set of surface area histogram bins represented by the descriptors: **PIP1, PIP2, PIP3, PIP4, PIP5, PIP6, PIP7, PIP8, PIP9, PIP10, PIP11, PIP12, PIP13, PIP14, PIP15, PIP16, PIP17, PIP18, PIP19,** and **PIP20**. The PIP bins extend from 0.420 au to 0.780 au in increments of 0.02 au. PIP extrema do not appear in many of the best models that we have yet encountered, even though **PIPMin** corresponds to the important "I-bar" Politzer parameter. The Middle Lower (5-8) and Middle Upper (12-16) PIP histogram range appear in many diverse models of disparate phenomena. It seems as if the PIP parameters are correlated with a number of intermolecular binding modes, not the least of which is induced-dipole interactions. There is reason to believe that the PIP descriptors also carry information about the "hardness" or "softness" of a region of electron density, as well as Donor/Acceptor information. Quite frequently, PIP descriptors show up with Del(Rho)N and SIK parameters to describe differential solubility or hydrophobic / hydrophilic interaction tendencies.

Only the "New" (C and D Series) Regional and Molecular fields carry the descriptors discussed below. The first of this new class are: **BNP, BNPMin, BNPMax, BNPAvg, BNP1, BNP2, BNP3, BNP4, BNP5, BNP6, BNP7, BNP8, BNP9,** and **BNP10**. **BNP** stands for "Bare Nuclear Potential", a reflection of the fact that this is the quantity which is being mapped on the the molecular electron density isosurface. While it may seem unusual at first, the value of the **BNP** does not vary much over the Van der Waals surface of a molecule. This is actually another way of saying that the isodensity-derived Van der Waals surface locus is almost obtainable through generating an appropriate **BNP** isosurface. Of course, one might argue that there are no electrons present in such a model, so how could it have any information about the electron density? The fact is that the geometry and orientation of the nuclei are, in fact, a consequence of the (now missing) electron density, and the strength of the **BNP** field when mapped onto a genuine electron density isosurface will give information about where the nuclear - electron attractive forces and the electron density repulsion forces are in or out of balance. This provides complementary information to that obtained through electrostatic potential information. As the Laplacian of the density represents a balance between the G and K electronic kinetic energy densities (one positive, one negative), the Electrostatic Potential represents a balance between the nuclear-electron attraction and the electron-electron repulsion forces. The overall surface-integrated value, the integral average of this field and the extrema of this property field are all used as descriptors: **BNP, BNPMin, BNPMax,** and **BNPAvg**. There are ten histogram fields for this property as well: **BNP1, BNP2, BNP3, BNP4, BNP5, BNP6, BNP7, BNP8, BNP9,** and **BNP10**. This descriptor is so new that there are no trends concerning its use. It would make sense that the **BNP** family of descriptors would work well with the **EP** and **PIP** parameters to describe polar interactions and hydrogen bonding. Other uses will have to be empirically discovered.

Another new addition to Molecular and Regional QMF files are the Fukui Radical Reactivity Indices. **Fuk, FukMin, FukMax, FukAvg, Fuk1, Fuk2, Fuk3, Fuk4, Fuk5, Fuk6, Fuk7, Fuk8, Fuk9, Fuk10**. This set of indices was added after the October 1996 Kodak Meeting in which there was emphasis on the radical decay process of film products and the stabilizers added to help with film longevity. The Fukui indices are related in a way to the **PIP** indices, in that they both involve a perturbation expression which is meant to describe the spatial distribution of radical reactivity. In the **PIP** case, the molecular surface is encoded with energy-weighted orbital densities, while in the Fukui case, there is a selectable denominator term which places the reactivity index on a cationic, radical or anionic scale. The Fukui index, as implemented here, describes radical reactivity. More data will have to be collected in order to form a more coherent model of **Fuk** descriptor behavior.

Another new Regional and Molecular index set to be added are the Laplacian-derived indices: **Lapl, LaplMin, LaplMax, LaplAvg, Lapl1, Lapl2, Lapl3, Lapl4, Lapl5, Lapl6, Lapl7, Lapl8, Lapl9,** and **Lapl10**. As stated in the earlier discussion of the **G** and **K** electronic kinetic energy density balance, the Laplacian has been implicated as a descriptor in electrophilic aromatic substitution partial rate factors and in Donor/ Acceptor interactions. The Laplacian (or Del^2(Rho)) is the trace of the second derivative matrix of the electron density at any point in space. The truly indicative Laplacian phenomena are the negative peaks which form near the outer core regions of the electron density of molecules. These peaks, when in regions of non-bonded electron density or in regions of electrophilic attack, tend to be able to predict the rates of these reactions by their magnitude. Since the parameters that are derived from the Laplacian of the density are far from these negative peak regions, the "Laplacian Shadows" of these peaks on the Molecular Van der Waals surface are the best indicators of what is going on internal to the molecular surface. There is not enough information to comment on the use of the **Lapl** class of data, but there is reason to believe that a combination of **Lapl**, **Fuk** and **BNP** indices will be able to address the radical decay processes of film products and stabilizers.

The **Abs** type of descriptor utilizes the same raw properties that are used in the surface property-weighted property indicators described above. The main difference between **Abs** and non-**Abs** descriptors is that the former type can be much more easily interpreted graphically by examining molecular surface property graphics. **Abs**-type descriptors are also much less dependent upon the grid size chosen for the surface polygons as well as the orientation of the molecule in space.