In the centrosymmetric zone, $R_K$ and $R_C$ differ only by a normalizing denominator:

$$R_K = R_C \cdot \frac{\sum_{hkl} |\Delta F|}{\sum_{hkl} |F_{PH}|}.$$

Kraut $R_K$ values were calculated in the work to follow. These can be converted roughly to $R_C$'s by multiplying by 3·3 for Pt, 5·0 for Hg, and 3·6 for the double derivative.

The 'four-data' check sign set was the result of the combination of the Pt30w data, another 1:1 Pt/cytochrome mole ratio set photographed at six weeks (Pt6w), and Hg photographed at one week (Hglw) and again at fourteen weeks (Hg14w). Scale factors, $x$ and $y$, coordinates and degrees of substitution, $A$, were refined until no further significant changes occurred. The duplicate heavy atoms refined to the same sites to within 0·05 Å but differed in degree of substitution. Figures of merit and Kraut $R_K$ values are listed in Table 4. The 'five-data' set of Table 4 consists of the above four and the double derivative.

Difference Fourier maps were calculated with individual terms both unweighted and weighted with the figure of merit for each reflection. No practical advantage was seen in the weighted maps in terms of added information or detail; moreover these maps had the disadvantage of having peak heights reduced by a factor of around 0·6 (the mean figure of merit). All maps shown in this paper are unweighted.

A complete series of cross difference maps was calculated, using parent protein signs determined from one derivative and $\Delta F$ values from a different derivative. In this context, the expression 'mercury signs' is to be interpreted as meaning the signs of the parent protein obtained by using the mercury derivative data alone, and not the signs of the mercury atom contribution itself. Crossover terms, where $F_{PH}$ and $F_P$ have opposite signs and where the heavy atom contribution

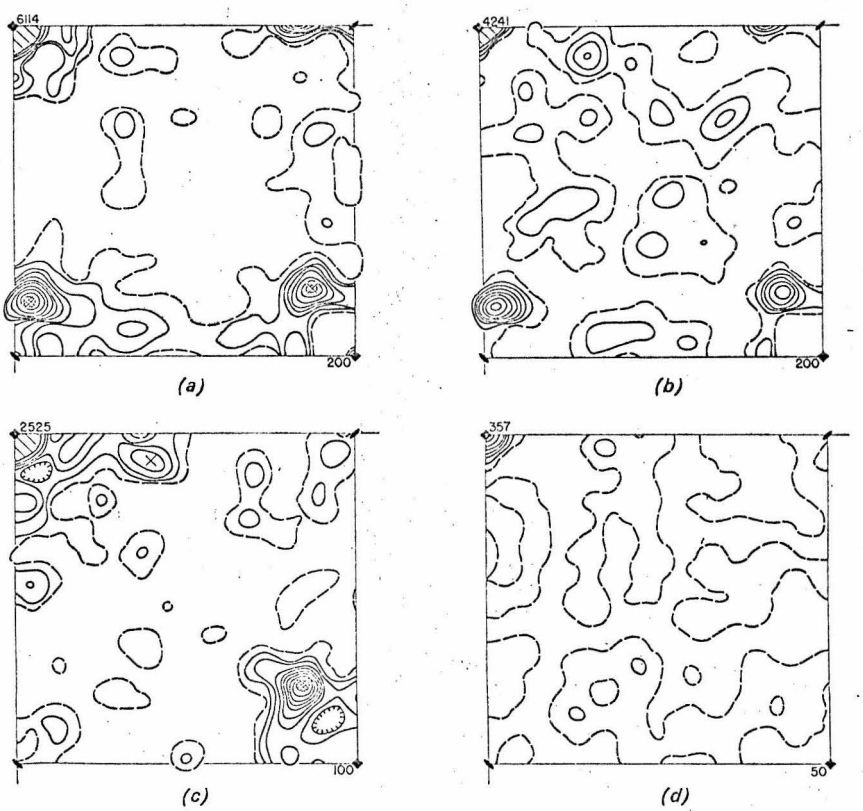



Fig. 1. $(\Delta F)^2$ difference Patterson maps of horse heart cytochrome C in $hk0$ projection at 3 Å resolution. All maps are to the same arbitrary scale. Contour intervals are marked in the lower right corner and the height of the origin peak in the upper left. The zero contour is dashed. Coordinates $u$ (horizontal) and $v$ (vertical) run from 0 to ½, with the origin in the upper left corner. (*a*) $PtCl_4^{2-}$ derivative (Pt30w). Interpreted as arising from a single Pt site per molecule. The expected single peak is marked by × and the double peak by #. (*b*) $PtCl_4^{2-}$ map as in (*a*) but with the innermost 20 of the 180 reflections omitted. Note the essential similarity to (*a*). (*c*) Mersalyl derivative (Hg14w). Single site, peaks indicated as in (*a*). (*d*) 'Null Patterson' map comparing two unrelated sets of parent cytochrome data to give some idea of the experimental noise level.

is equal in magnitude to the *sum* of the magnitudes of $F_{PH}$ and $F_P$, were rare and unimportant, and were neglected in the difference maps.

### Valid derivatives: platinum and mercury

Fig. 1 shows the $(\Delta F)^2$ difference Patterson projections down the *c* axis for Pt30w, Hg14w and for one set of parent cytochrome data against another independent parent set. This last map, Fig. 1(*d*), provides an estimate of the amount of background noise to be expected from errors in intensity data, and suggests that features over 50 to 100 in the other maps are probably to be ascribed to genuine structural changes between parent and derivative and not to experimental data error. The double peak at $(x+y, y-x)$ is plain in both Pt and Hg maps, as is the single peak at $(2x, 2y)$. Initial mean figures of merit (Table 4) from phase determination were around 0·4 for each of the derivatives taken alone, and 0·46 for the combination.

$\Delta F$ difference Fourier maps of three types for Pt and Hg are shown in Figs. 2 and 3: self-sign maps using signs determined by the derivative in question [Figs. 2(*a*) and 3(*a*)], cross-sign maps using signs determined by the *other* derivative [Figs. 2(*b*) and 3(*b*)], and check maps using refined four-data signs [Figs. 2(*c*) and 3(*c*)]. The self-sign maps give the best reproductions of the assumed peaks, of course, but prove nothing. Feedback of the initial assumptions in such cases is at its strongest, and any peak chosen at random for sign determination would have reappeared on the map and given the appearance of being 'confirmed' (see Fig. 9(*a*), for example). The real confirmations of the two derivatives are their cross-sign maps. There is no reason for a peak to show up where it does in Fig. 2(*b*), for example, unless: (*a*) the $\Delta F$'s are meaningful because of the correctness of the Pt derivative, and (*b*) the signs are meaningful because of the correctness of the Hg derivative. The self-sign Pt map, because of its bias in favor of a single-site Pt model, actually obscures information. The improvement in $R_K$ for Pt30w in Table 4 from 17·5 to 13·4% is mainly a result of the addition of a half-weight secondary site immediately to the right of the principal site [refined position at × in Fig. 2(*a*)]. The asymmetry of the Pt peak is most pronounced in the cross-sign map, and the secondary site is less marked relative to the primary site in just that map whose signs are most heavily influenced by a one-site model.

The maps obtained using the two-derivative sign set and the refined four-data sign set (Figs. 2(*c*) and 3(*c*), and Table 3, *A*) offer no proof of derivative correctness beyond that of the cross-sign maps in spite of their improved appearance. This improved appearance arises from the introduction into the sign analysis of just those sites whose validity is being tested, plus the further advantage of least-squares fitting of the assumed models to the data. It is no surprise that they look superficially better. It is true that the least-squares program will often refine the effective substitution number of an incorrect site to zero (see erronium trials, below, and also Dickerson & Palmer, 1967). But it cannot add an omitted site, and the use of refined full-derivative signs clouds the proof by interrelating what would otherwise be two totally independent demonstrations of correctness.

The only $\Delta F$ difference maps which are really worth anything in proving out derivatives, in summary, are the cross-sign maps in which the sign set used has no dependence upon the parameters of the derivative whose validity is being tested. Self-sign maps are utterly worthless except as journal illustrations.

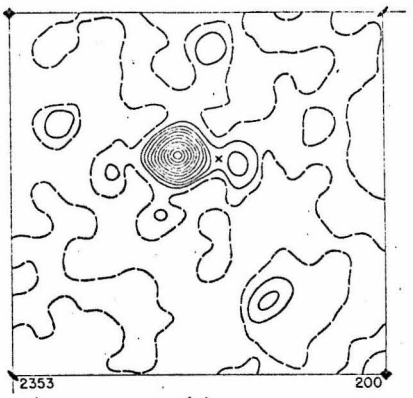
### An example of an incorrect interpretation: $PdCl_4^{2-}$

Fig. 4(*a*) shows the difference Patterson map between crystals in 4·3 *M* mixed phosphate buffer and similar crystals with $PdCl_4^{2-}$ added. A study of the effect of medium on $|\Delta F_{hk0}|$ in ammonium sulfate and in 4·3 and 5·0 *M* phosphate buffer showed that, although several of the low-order terms changed in intensity, only the 100 and 110 reflections actually changed sign. This change of sign was accounted for in all the $\Delta F$ maps to follow. The 'interpretation' of the Pd map will illustrate some of the potential hazards of the process.

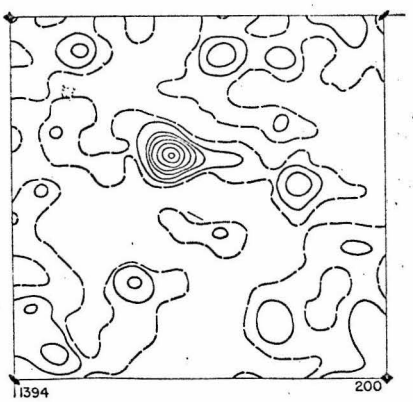
Although this Pd map is noisier than either of the Pt or Hg maps, it can be fitted by a two-site model. These two sites, marked *A* and *B* in Fig. 5, account for the peaks in Fig. 4(*a*) which are marked by ×, ≠, or #. Only the two satellite peaks below and to the upper right of the very large peak near the origin are unexplained. If the data are cut off at 6 Å resolution, then these peaks merge and the interpretation is even more convincing. At 4 Å the peak heights are very nearly correct; the highest peak is interpreted as the near-superposition of two double peaks, the next two as near-superpositions of a double and a single, and the next two as double sites. With due allowance for origin diffraction ripple, experimental error particularly in low-order reflections, and a certain degree of non-isomorphism, it is easy to be persuaded that the interpretation is right.

The Pd self-sign difference Fourier map [Fig. 6(*a*)] supports this conclusion and even suggests a minor third site, *C*. With two exceptions, all ten of the new *C*–*C*, *C*–*A* and *C*–*B* vectors required do show up on or near features of the Patterson map. Nevertheless, this interpretation is wrong from start to finish, and it is important to see what kind of warning signs reveal this.

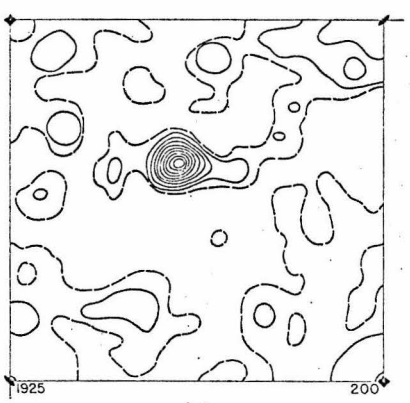
The first sign of trouble is the low figure of merit for the Pd derivative taken alone, 0·3 as compared with 0·4 for Pt or Hg. Admittedly, the absolute value of the figure of merit is a shaky measure of correctness. Experiments with cytochrome C and with triclinic hen egg-white lysozyme suggest that the mean figure of merit rises sharply with (*a*) the number of derivatives, (*b*) the number of sites per derivative, (*c*) the number




(a)

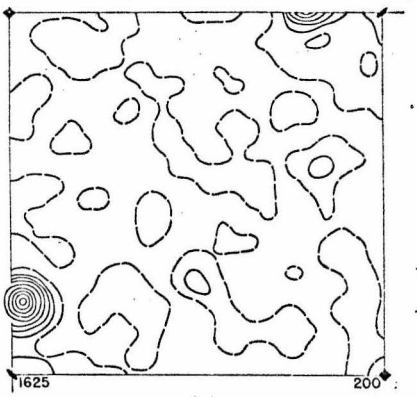



(b)

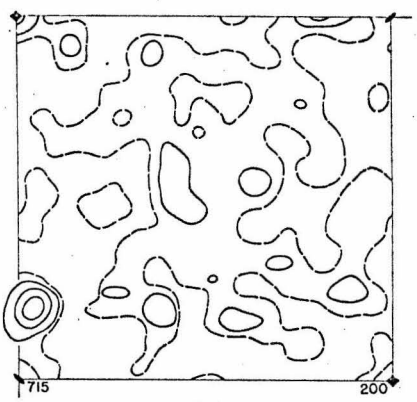



(c)

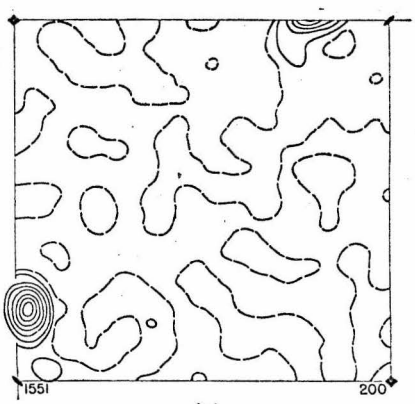
Fig. 2. $\Delta F$ Difference electron density projections using platinum $\Delta F$'s and the sign sets indicated. These maps and all $\Delta F$ maps to follow are on the same arbitrary scale. The contour interval in each map is given in the lower right corner and the height of the main peak in the lower left. Zero contours are dashed. Coordinates $x$ (horizontal) and $y$ (vertical) run from 0 to $\frac{1}{2}$, with the origin in the upper left corner. All maps are unweighted. (*a*) Pt signs. (*b*) Hg signs. (*c*) 4-Data signs.




(a)



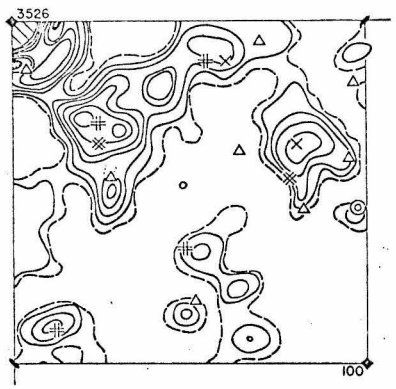

(b)



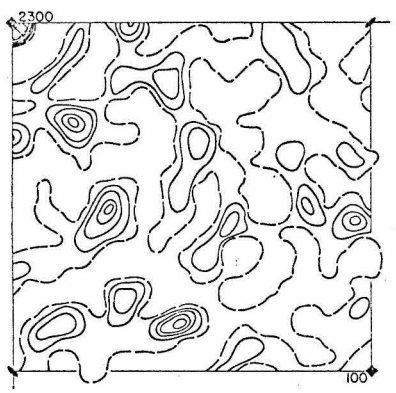

(c)

Fig. 3. $\Delta F$ difference electron density projections using mercury $\Delta F$'s and the sign sets indicated. Same conventions and scale as Fig. 2. (*a*) Hg signs. (*b*) Pt signs. (*c*) 4-Data signs.

of parameters adjusted by least squares per site (such as anisotropic temperature factors), and (*d*) the extent to which the r.m.s. errors in the derivatives are underestimated. As an example, the mean figure of merit for a three-dimensional phase analysis at 6 Å of triclinic hen egg-white lysozyme using three derivatives with 5, 6, and 6 sites was 0·83 (Dickerson & Steinrauf, unpublished), yet the analysis is now believed to have been at least partially in error. In a variant of this particular test, when the same lysozyme data were used, with the same number of sites in the three derivatives, but with atomic coordinates chosen from a table of random numbers, the mean figure of merit was 0·69. But in spite of the problems in interpreting absolute values of *m*, the *relative* values and changes in mean




(*a*)




(*b*)

Fig. 4. $(\Delta F)^2$ Difference Patterson maps of the $PdCl_4^{2-}$ derivative in 4·3 *M* phosphate buffer against parent crystals in 4·3 *M* phosphate buffer. Same conventions and scale as Fig. 1. (*a*) Full 4 Å data. Expected peaks are marked as follows:

× # Single and double peaks from sites *A* and *B* taken individually as shown in Fig. 5.
≠ Cross vector peaks between sites *A* and *B*.
○ ◎ Single and double peaks from site *C* of Fig. 5.
△ Cross vector peaks between site *C* and site *A* or *B*.

(*b*) Same as (*a*), but with the innermost 20 reflections omitted. Compare the destruction of features of (*a*) with the analogous Pt maps [Fig. 1(*a*) and 1(*b*)].

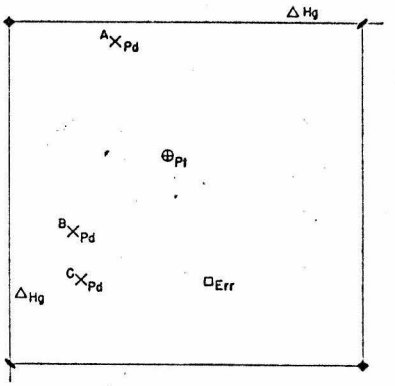
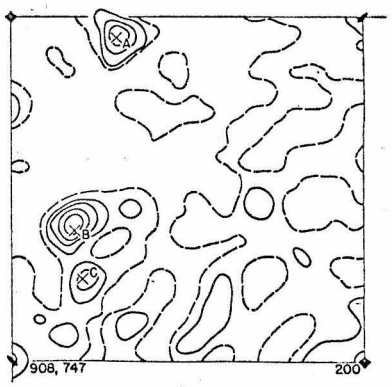
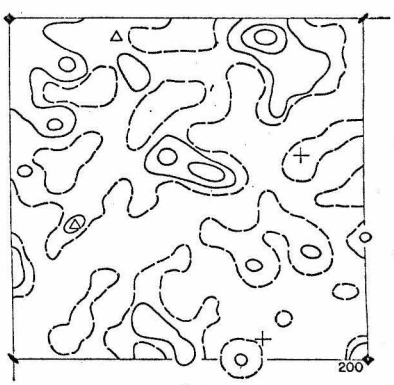



Fig. 5. Heavy atom sites. Same axis conventions as for $\Delta F$ maps. Sites are: ⊕ Pt; △ Hg; × Pd (3 sites); □ Erronium.




(*a*)




(*b*)

Fig. 6. $\Delta F$ Difference electron density projections using Pd $\Delta F$'s and the sign sets indicated. Same conventions as Fig. 2. (*a*) Pd signs. (*b*) 4-Data signs. The two principal sites are marked by △, and their symmetry-related alternates by +.

figure of merit during the course of analysis are useful in differentiating between good and bad derivatives under controlled conditions.

The clearest indication of trouble comes from the $\Delta F$ maps. The Pd difference map with four-data signs [Fig. 6(*b*)] fails to reproduce the Pd sites, either as originally chosen or with a shift of origin by $(\frac{1}{2},\frac{1}{2},0)$. Instead, 'ghost' Pt and Hg peaks appear, reflecting the fact that, although the sign set is a reasonably correct parent protein set, it has been biased in favor of the Pt and Hg sites used to obtain it. In the corresponding map using Pt signs alone, the Pt ghost peak becomes stronger at the expense of the Hg, and with Hg signs the reverse is true. The Pt and Hg difference maps using Pd signs (Fig. 7) fail to reproduce the Pt and Hg sites, and the Pt map appears to have ghost Pd peaks as well.

In addition to these direct space tests, a test in reciprocal space of agreement between $hk0$ signs determined by Pd and those from Pt, Hg, or any combination of them, demonstrates the incompatibility of the Pd set. A symmetry-permissible shift of origin in Pd by $(\frac{1}{2},\frac{1}{2},0)$, resulting in the reversal of sign of all reflections with $h+k$ odd, is of no help, and the conclusion must be drawn that either the Pd or the Pt and Hg taken together are wrong.

The Pd trials illustrate one of the pitfalls of interpreting difference Patterson maps, namely the ability to fit anything with enough concentrated effort. With two assumed sites per molecule, in space group $P4_1$, there will be six double peaks and two single peaks in the quadrant shown. With three sites per molecule, there will be three single peaks and fifteen double peaks. With such a great number of peaks available and with the low order 'checkerboarding' effect discussed below, some kind of a fit or another is assured.

At low resolution, and especially in projection, errors in a handful of low-order reflections can introduce serious error ripples into the difference Patterson map. If only one or two terms are involved, the erring reflections can be identified, but if several are acting in concert, it may not be apparent that low-order error fringes are the cause of the observed peaks. Errors in

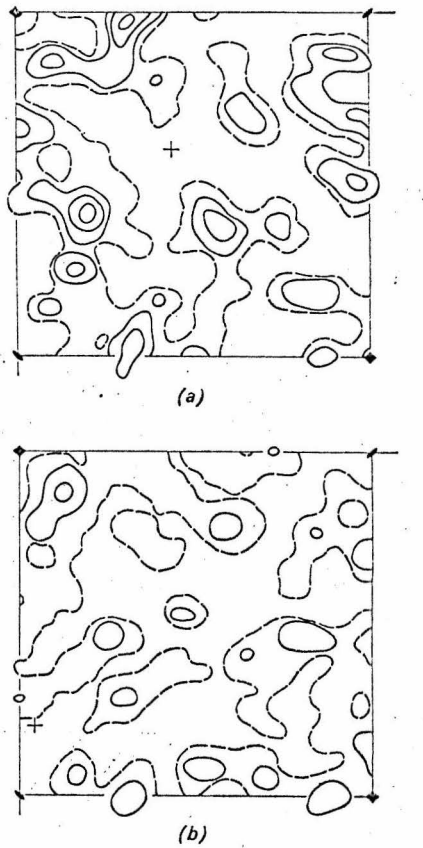


Fig. 7. (*a*) Pt and (*b*) Hg $\Delta F$ difference maps with Pd signs. Expected heavy atom sites are marked with a +.

Table 3. *Peak heights in ΔF difference electron density maps*

*A.* Pt30w difference maps

| Signs: | Pt | Hg | Pt+Hg | 4-Data | Pd | Err |
|---|---|---|---|---|---|---|
| Pt peak heights: | 2353 | 1394 | 2067 | 1925 | (−700)† | (0)† |
| Heights with Err*: | 2139 | 1231 | 2134 | — | — | — |

*B.* Hg14w difference maps

| Signs: | Pt | Hg | Pt+Hg | 4-Data | Pd | Err |
|---|---|---|---|---|---|---|
| Hg peak heights: | 715 | 1625 | 1553 | 1551 | (−50)† | (142)† |
| Heights with Err*: | 919 | 1425 | 1423 | — | — | — |

*C.* Erronium difference maps

| Signs: | Err | Err+Pt | Err+Hg | Err, Pt, Hg | 4-Data |
|---|---|---|---|---|---|
| Err peak heights: | 1234 | 700 | 677 | (345)† | (−16)† |

*D.* Double derivative difference maps

| | Unweighted maps | | | Weighted with figure of merit | | |
|---|---|---|---|---|---|---|
| Signs: | Pt | 4-Data | Hg | Pt | 4-Data | Hg |
| Pt peak heights: | 1711 | 1642 | 1232 | 1144 | 1367 | 643 |
| Hg peak heights: | 283 | 722 | 803 | 167 | 603 | 562 |
| Pt/Hg ratio: | 6·05 | 2·28 | 1·54 | 6·85 | 2·27 | 1·14 |

* Peak heights obtained from sign sets after erronium is added to the analysis.
† No longer the highest feature on the difference map.

somewhat higher order terms in tetragonal space groups lead to the grid pattern of fringes which Kraut calls 'checkerboarding'.

If it proves possible to fit an eight-atom model – two per molecule – to $8\times7=56$ peaks in a three-dimensional difference Patterson map, the temptation to accept it as valid is strong. But the danger lies in the fact that the nodes or intersections of a set of low-order error fringes *themselves* will form a vector set. In the presence of such features, one will automatically be able to find four, or six, or eight atomic sites which will explain peaks which are really nothing more than low-order fringe nodes. However, an atomic site set obtained in such a way will itself tend to show the regularity of the error nodes, with sites falling in parallelograms or on a small number of parallel lines. This may be the proper explanation of the difference Patterson map of the $HgI_4^{2-}$ derivative of triclinic lysozyme shown in Fig. 10 of Dickerson (1964).

The simplest way to see if the inner reflections are producing fake detail is to leave them out and see what happens to the map. The Pt and Pd difference Patterson maps of Figs. 1(*b*) and 4(*b*) differ from those of Figs. 1(*a*) and 4(*a*) only in having the innermost 20 of the 180 reflections removed – all those within the 11·5 Å limit. The features of the Pd map are now obliterated, leaving nothing but noise. The Pt map is virtually unaffected. As the null Patterson of Fig. 1(*d*) indicates that the features of the Pd map are higher than the experimental noise level, the peaks of Fig. 4(*b*) most probably arise from extensive non-specific or at least multi-specific binding to the protein, and the derivative is useless.

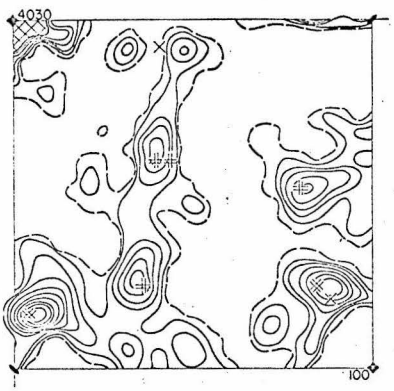



Fig. 8. $(\Delta F)_2^2$ Patterson map of the double derivative with Pt and Hg diffused in simultaneously. Pt–Pt and Hg–Hg single and double vectors are indicated by × and #, Pt–Hg double weight cross vectors by ≠.

### A valuable control: the double derivative

The best check on the validity of two individual derivatives is the simultaneous introduction of both sites into the same crystal. Now, in addition to the self peaks, there must appear the *cross* peaks in the difference Patterson map. It is difficult to see how a fortuitous combination of error ripples of the type discussed in the previous section could produce just the required peaks in the individual maps, and then these peaks plus just the right extra peaks in the double derivative Patterson map. Fig. 8 shows such a difference Patterson map for the simultaneous Pt/Hg double derivative. The cross peaks (#) show up exactly where they are predicted to be, and the assumed model explains all of the highest features of the map.

An example of the greater sensitivity of the weighted maps to feedback errors is provided by the Pt/Hg peak height ratios in difference Fourier maps of the double derivative using sign sets from various sources (Table 3, *D*). When the sign set has been obtained from a single derivative corresponding to one of the two sites, that site is emphasized in the difference map at the expense of the other, leading to peak ratios in the unweighted maps ranging between 6·05 and 1·54. The final refined value was 2·8, close to that produced by the four-data sign set. The weighted maps, although producing the same nearly correct value with the four-data signs, exaggerate the emphasis on the sign-determining site when single-derivative signs are used.

Table 4. *Reciprocal space test of derivative quality*

| | | $R_K$ | | | | | |
|---|---|---|---|---|---|---|---|
| Derivatives | *m* | Pt6w | Pt30w | Hg1w | Hg14w | Double | Other |
| Pt30w, unrefined | 0·370 | | 17·0 | | | | |
| Hg14w, unrefined | 0·396 | | | | 10·6 | | |
| Pt+Hg, unrefined | 0·459 | | 19·0 | | 12·4 | | |
| Four-data, refined | 0·575 | 15·5 | 17·5 | 11·3 | 11·1 | | |
| Double derivative, unrefined | 0·328 | | | | | 17·0 | |
| Five-data, refined, with secondary Pt site | 0·641 | 12·6 | 13·4 | 11·1 | 10·7 | 13·8 | |
| Palladium unrefined | 0·291 | | | | | | 15·3 |
| Erronium, unrefined | 0·272 | | | | | | 17·4 |
| Err+Pt | 0·360 | | 19·2 | | | | 22·9 |
| Err+Hg | 0·361 | | | | 12·3 | | 23·3 |
| Err+Pt+Hg, unrefined | 0·430 | | 19·9 | | 12·9 | | 26·5 |
| Err+Pt+Hg, refined | 0·485 | | 16·9 | | 10·8 | | 21·0 |

### The effect of a wrong derivative on subsequent analysis: erronium

The checks discussed above should ordinarily be sufficient to exclude a wrong interpretation. But what would be the effect of adding a totally erroneous derivative to the sign analysis? How serious would the feedback problem be, and at what point would the mistake be discovered?

To answer this question, such a site was selected and was used with the Pd data with the four innermost salt-sensitive reflections discarded. The site selected (Fig. 5) was chosen so as to be removed from any real site, not to be on any obvious low-order fringes which included a real site, and not to be compatible with the Pd difference Patterson map. This wrong site with its now pseudo-random $\Delta F$ data will be referred to as 'erronium', Err. Structure factor calculations, Wilson plots for $A$ and $B$, and sign determinations were carried out for Err just as for either of the true derivatives, and the parameters obtained are given in Table 2. The Wilson plot was more ragged than usual, but not disturbingly so.

The self-sign map [Fig. 9(*a*)] is reassuring; the peak is almost as prominent as is the Hg peak in its own self-sign map. But the addition of either Pt or Hg to the sign analysis halves the Err peak, and the addition of both good derivatives wipes it out altogether [Fig. 9(*b*)–(*d*)]. In contrast, going from either a Pt or a Hg self-sign map to the two-derivative Pt+Hg map causes only a 5–10% drop in peak height. The map with Err (or Pd) $\Delta F$'s and the four-data signs [Fig. 6(*b*)] shows no trace of the Err peak.

The Pt and Hg cross-sign maps using Err signs show no trace of the expected peaks (Fig. 10). But the addition of Pt to Err in the sign analysis brings in the resultant Pt peak in the difference map almost as strongly as with Pt signs alone. In fact, each of the Pt maps with signs of Pt+Err, Hg+Err and Pt+Hg+Err is nearly as good as the corresponding map without Err (Table 3, *A*, *B*). Exactly the same behavior is observed with Hg $\Delta F$ maps and the corresponding sign sets. The Err contribution is so meaningless that to a good approximation the sign analysis behaves as if the Err is simply not there. It only serves as a perturbing source of random error which lowers the quality of the maps and raises background noise. The relatively small drop in mean figure of merit upon addition of erronium (Table 4) corroborates the unsystematic nature of its contribution.

The refinement program provides the most dramatic evidence of the incorrectness of erronium. Before re-

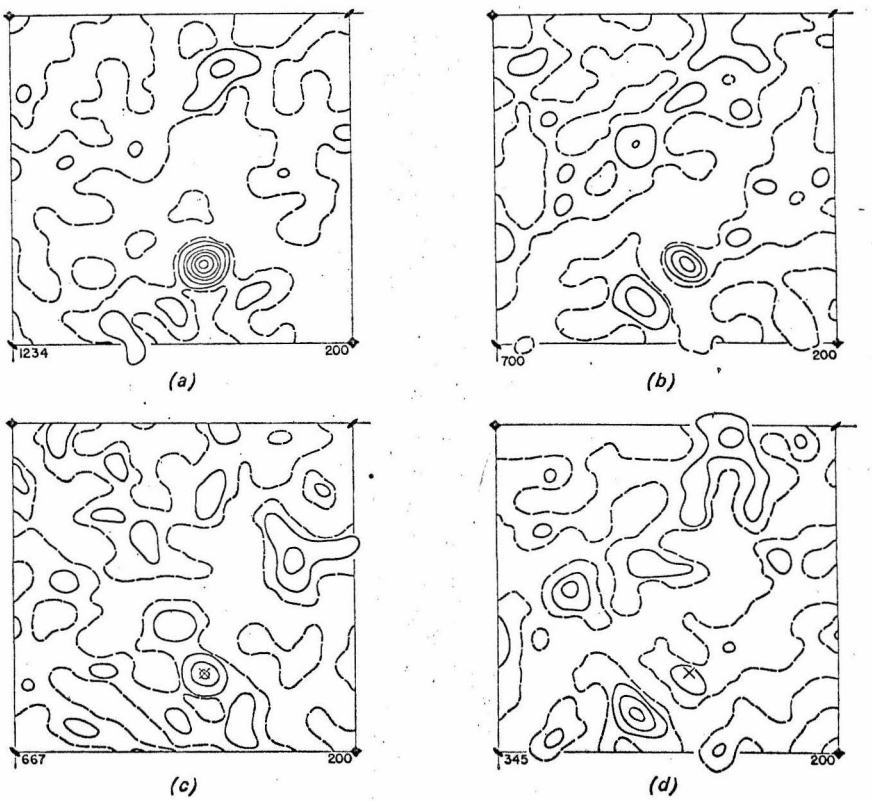



Fig. 9. Erronium $\Delta F$ maps with signs indicated. (*a*) Err signs. (*b*) Err+Pt signs. (*c*) Err+Hg signs. (*d*) Err+Pt+Hg signs.

finement, the four-data set gave a mean figure of merit of 0·52, and ten cycles of refinement brought this up to 0·58. The Err+Pt+Hg set before refinement gave a value of 0·43, and ten cycles brought this up to 0·48. However, in the course of this refinement the occupancy number of the erronium site fell drastically as shown in Fig. 11.

Thus, in this example of a bad derivative, the bad derivative difference map was damaged by the introduction of only one good derivative and was wiped out by two; the bad sign set could not pull in the good derivatives in difference maps, and the bad site refined down towards zero occupancy when permitted to do so.

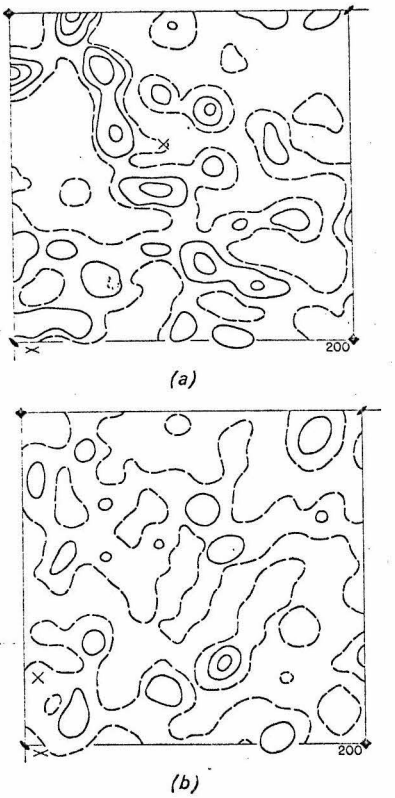



Fig. 10. (*a*) Pt and (*b*) Hg $\Delta F$ difference maps with Err signs. Expected heavy atom sites are marked with a ×.

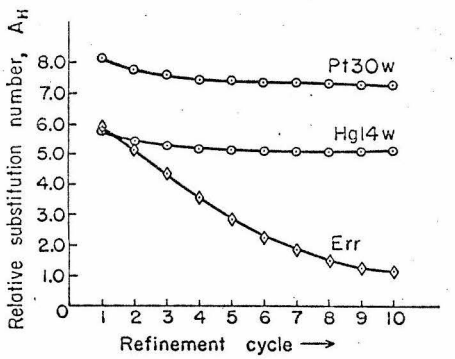



Fig. 11. Effective atomic numbers or site occupancy numbers of Pt, Hg, and Err during phase refinement.

## Conclusions

These trials seem to illustrate some fairly general principles:

(1) In interpreting difference Patterson maps, there is considerable danger of deception by false detail from errors in low-order terms, particularly at 6 Å resolution. If many sites are required, if vector peaks from these sites tend to overlap in multiple peaks, if the sites themselves tend to fall on a common plane or to form parallelograms or anything like them, and especially if the Patterson map is sensitive to the removal of any selection of inner reflections, then the interpretation should be viewed with mistrust.

(2) Difference Patterson maps of multiple-site derivatives (3 or more per asymmetric unit in $P4_1$, for example) probably will not be interpretable in isolation and will have to be brought in with the aid of other derivatives. But this roundabout interpretation must then be shown to be compatible with the original difference Patterson map.

(3) The surest proof of the validity of a derivative by difference Fourier methods is a cross-sign map using signs or phases which are absolutely untainted by information about the derivative in question.

(4) A completely wrong derivative will have less effect on the phase analysis than might have been expected, especially in the presence of other very good derivatives. Careful use of cross difference Fourier maps will reveal the presence of a spurious derivative.

(5) Trends in the refinement of the substitution numbers will often point out false derivatives or false secondary sites. Although the absolute values of the mean figure of merit can be deceptive, relative changes in $m$ as derivatives are combined *are* informative.

(6) The feedback problem need not be severe. Ghost peaks, while sometimes visible, would seldom be confused with real peaks unless one were overenthusiastic about choosing 'subsidiary sites'. If the minor sites of one derivative turn out to coincide with the principal sites of other derivatives, then there are grounds for skepticism.

An investigator who publishes a low-resolution protein structure analysis is asking the reader to take an unusually large amount on faith. The results, once obtained, do not necessarily make any obvious chemical sense, the usual way of casually judging a structure analysis. As a compensation, therefore, the path of analysis should be crystal-clear.

A reasonable minimum standard of publication might include the following:

(*a*) Complete $F$ data for parent and derivatives. At low resolution, this amounts to only 1000–1500 reflections per derivative, a not excessive number even by

conventional standards. Since these data are the justification for everything else, they should be easily accessible to the reader.

(*b*) The $(\Delta F)^2$ difference Patterson map for *each* derivative used, whether it was interpretable by itself or not. The map should have marked on it the locations of the vectors expected from the heavy atom sites as finally adopted.

(*c*) A summary of the manner in which the heavy atom sites were deciphered, and in particular how those derivatives whose difference Patterson maps were uninterpretable in isolation were pulled into the analysis.

(*d*) $\Delta F$ difference maps for each derivative using phases or signs obtained from other unrelated derivatives. The parameters of these derivatives should be as they were before any refinement in combination with the derivative in question.

(*e*) Mean figures of merit and other refinement criteria such as the Kraut $R$ factor for each derivative separately (or for pairs of derivatives in three dimensions), and for the final combination of all derivatives before and after refinement.

(*f*) Commentary on any unusual features of the refinement, such as the previously mentioned wiping out of the erronium site, which would permit one to judge the derivatives.

The opportunities for self-deception in a low resolution analysis are limitless. If the fundamental difference Patterson maps are interpretable, then their publication should be a matter of record. If some are *not* interpretable, and if the derivatives are used in the phase analysis, then publication of the maps becomes a matter of obligation. Enough supplementary information should then be provided to convince the average crystallographer that the derivatives are valid *in spite of* the uninterpretability of the difference Patterson maps. In view of the difficulties of interpreting the final structure, the onlooker may legitimately ask, 'If you don't know where you are going, how do you know when you are there?' The only answer is that the course of analysis must be so transparent and so obvious that *any* end product, no matter how unexpected in appearance, will be accepted.

We would like to thank Drs Jon Bordner and David Eisenberg for their help in the calculations involving the double derivative, and Miss Lillian Casler for the preparation of all of the figures. The authors are indebted to the United States Public Health Service for their support in the form of research grant GM 12121, under which this work was carried out. One of the authors (J.W.) is also the holder of a National Institutes of Health predoctoral traineeship.

### References

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794.

CULLIS, A. F., MUIRHEAD, H., PERUTZ, M. F., ROSSMANN, M. G. & NORTH, A. C. T. (1961). *Proc. Roy. Soc.* A, **265**, 15.

DICKERSON, R. E. (1964). In *The Proteins*, Vol. II, p. 623. New York: Academic Press.

DICKERSON, R. E., KENDREW, J. C. & STRANDBERG, B. E. (1961*a*). *Acta Cryst.* **14**, 1188.

DICKERSON, R. E., KENDREW, J. C. & STRANDBERG, B. E. (1961*b*). In *Computing Methods and the Phase Problem in X-ray Crystal Analysis*, p. 236. New York: Pergamon Press.

DICKERSON, R. E. & PALMER, R. A. (1967). *Acta Cryst.* Submitted for publication.

KRAUT, J. (1961). Private communication.

KRAUT, J., SIEKER, L. C., HIGH, D. F. & FREER, S. T. (1962). *Proc. Nat. Acad. Sci. Wash.* **48**, 1417.

LIPSCOMB, W. N., COPPOLA, J. C., HARTSUCK, J. A., LUDWIG, M. L., MUIRHEAD, H., SEARL, J. & STEITZ, T. A. (1966). *J. Mol. Biol.* **19**, 423.

MARGOLIASH, E. (1967). In *Methods in Enzymology*. (Edited by S. P. Colowick and N. O. Kaplan). In the press.

MUIRHEAD, H. (1966). Private communication.

*Printed in Denmark by P. J. Schmidts Bogtrykkeri, Vojens*

*J. Mol. Biol.* (1967) **29**, 77–95

# A Centrosymmetric Projection at 4 Å of Horse Heart Oxidized Cytochrome *c*

RICHARD E. DICKERSON, MARY L. KOPKA, CHARLES L. BORDERS, JR., JOAN VARNUM, JON E. WEINZIERL

*Gates and Crellin Laboratories of Chemistry*
*California Institute of Technology, Pasadena, California, U.S.A.*

AND

EMANUEL MARGOLIASH

*Division of Molecular Biology,*
*Abbott Laboratories, North Chicago, Illinois, U.S.A.*

(*Received 8 May 1967*)

Horse heart ferricytochrome *c* has been crystallized in space group $P4_1$ with one molecule per asymmetric unit and cell dimensions: $a = b = 58{\cdot}45$ Å, $c = 42{\cdot}34$ Å. A derivative search concentrating on transition metal complexes, organomercurials and labeled carboxymethylation reagents has produced two isomorphous single-site derivatives, $PtCl_4^{2-}$ and mersalyl. Data from the *hk*0 zone out to a resolution of 4 Å have been collected, heavy metal sites found and phases found and refined. The refinement behavior of mean figure of merit and Kraut *R* factor has been compared. The *hk*0 projection has been calculated and interpreted, and the interpretation checked with salt difference studies. The molecule appears to be a sphere of 31 Å diameter, with a center of packed hydrophobic side chains, a polypeptide chain framework and an outer covering of packed hydrophilic side groups.

## 1. Introduction

Cytochrome *c* has the dubious distinction at the moment of being the only cytochrome member of the mitochondrial terminal oxidation chain which can be isolated and characterized on the molecular level without loss of function. It is present in all aerobic organisms, and must have evolved very early in the history of life on this planet. The primary amino acid sequences of cytochrome *c* from a great many species are known (Margoliash & Schejter, 1966), and comparison of them has led to some very fruitful speculations on evolution at the molecular level.

Out of this comparative sequence work has come the observation that some regions of the chain, being almost totally invariant across the entire range of species, must be absolutely essential for the functioning of the molecule. Other regions, although constant for any one species, vary widely from one species to another. Hydrophobic residues tend to cluster, as do the great number of basic groups (24 out of 104 in horse), and to a certain extent hydrophobic and basic regions coincide. It is a matter of some interest to know how these different regions of sequence are arranged relative to one another in space. The manner of binding of the heme to the polypeptide chain is also of interest, as it is known from chemical studies to be different from that in myoglobin or hemoglobin. There is also evidence of at least a small change in conforma-

tion between ferri- and ferrocytochrome *c* (summarized in Margoliash & Schejter, 1966), so that for an understanding of the mechanism of action it becomes important to know the structure of both the oxidized and reduced forms.

As a first step towards this goal, an X-ray diffraction crystal structure analysis has been begun on the oxidized form of horse heart cytochrome *c*, and the results of a two-dimensional analysis are reported here, at a resolution of 4 Å. The three-dimensional 4 Å analysis is in progress.

## 2. Materials and Methods

### (a) *Extraction and crystallization*

Cytochrome *c* is extracted from frozen horse hearts by the method of Margoliash & Walasek (1967), using aluminum sulfate at pH 4·5 rather than trichloroacetic acid or sulfuric acid. Under these gentler conditions, a product is obtained which crystallizes from 90 to 95% saturated ammonium sulfate, 0·5 to 1·0 M in sodium chloride, at approximately pH 6·2, in a matter of days or weeks. The crystals are stable in the range of pH 4·5 to 8·0. Approximately 3 months are required for crystals in a new batch to grow to the size desirable for X-ray work (about 0·2 × 0·2 × 0·6 mm). A typical crystal preparation is shown in Plate I.

### (b) *Crystal data*

The protein crystallizes in tetragonal space group $P4_1$, as prisms of square cross-section elongated in the direction of the 4-fold, or *c*-axis, with length to width ratios of the order of 2 : 1 to 10 : 1. Only (100), (010) faces and faces of the {111} class are observed. Unit cell dimensions are: $a = b = 58{\cdot}45$ Å, $c = 42{\cdot}34$ Å. Photographs of the $h0l$ and $hk0$ zones are shown in Plate II. The crystal density as measured by flotation in chlorobenzene/bromobenzene mixtures is 1·264 g/cm$^3$. With the assumption of 4 molecules/unit cell or one/asymmetric unit, the crystal is then calculated to be only 44·9% protein by weight, the remainder being intermolecular liquid of crystallization. With a molecular weight of 12,400 and a unit cell volume of 144,700 Å$^3$, this crystal form has a vol./atomic mass unit of 2·92 Å$^3$/a.m.u., which is high in comparison with most proteins and even with other crystal forms of cytochrome *c* itself (Dickerson & Margoliash, manuscript in preparation). It is mathematically conceivable that there could be eight molecules per cell, in a crystal 90% protein by weight. But the intrinsic unlikelihood of this value when compared with other proteins, together with the mechanical softness of the cytochrome crystals, their susceptibility to heat disordering as observed by the fading out of the X-ray pattern above 80°F, their shrinkage upon drying and the fact that the diffused heavy-atom sites always occur 4 to a cell make such a tightly packed cell untenable.

### (c) *Preparation of derivatives*

The derivatives to be discussed in this paper were almost all prepared by diffusing heavy-atom groups into pre-grown crystals, the heavy-atom compounds having previously been dissolved in real or simulated mother liquor. In a typical preparation, 2 to 4 ml. of crystal suspension in 95% saturated ammonium sulfate 0·5 to 1·0 M in NaCl, was placed in a small plastic-lidded glass vial. A fresh, measured (usually near saturation) solution of $PtCl_4^{2-}$ or mersalyl in this ammonium sulfate–sodium chloride mixture was prepared and the pH adjusted to that of the crystal suspension with acetic acid or ammonium hydroxide. The correct volume of heavy-metal solution for the cytochrome to metal mole ratio desired was then pipetted into the glass vial and the solution gently swirled. The final pH was checked and adjusted if desired. The sample vials were stored at 5°C until used for photography. Total concentrations of cytochrome in the crystal suspensions before the addition of heavy-metal compounds were typically 0·6 to 1·8 μmoles/ml., and the added metal solution was typically 1/10 the volume of the crystal suspension. Metal to protein ratios varied from as much as 50 : 1 in some preliminary tests to as little as 1 : 1 in the final Pt crystals used for collection of three-dimensional data. In some of the early trials, the crystals tended to dissolve partially and then to regrow after addition of heavy-metal solution. In view of the quite open packing of the crystal structure, the distinction between diffusion and crystal growth in the presence of heavy-metal was felt

to be unimportant. All crystal preparations were stored at 5°C, but crystals were mounted and photographed at 20°C. No material improvement in picture quality was seen in crystals photographed at 5°C.

Three main types of heavy-atom derivative were tried: transition metal complexes, organomercurials and tagged carboxymethylation reagents. In addition, mixed phosphate buffer was investigated as a possible crystal medium. In all, 71 potential derivatives were tested in 110 different screening trials in ammonium sulfate by looking for intensity changes in 12·5° (4 Å) *h*0*l* precession camera photographs. Only two compounds produced usable changes: $PtCl_4^{2-}$ and mersalyl: HO–Hg–$CH_2$–CH(O–$CH_3$)–$CH_2$–NH–CO–(*o*–$C_6H_4$)–O–$CH_2$–COONa.

(i) *Transition metal complexes*

Our experience with diffused derivatives seems to have been in marked contrast to that of many other X-ray groups. Instead of finding many derivatives which produce intensity changes, and then encountering difficulty in interpreting the difference Patterson maps because of many sites or non-specific binding, we very rarely found intensity changes, but found them always associated with clean single-site binding (see below). This may be an effect of the abnormally high salt concentration and the difficulty of getting all but the most reactive heavy-atom groups into solution and near the protein at all. It may also be an effect of the competition of ammonia lone pair electrons with potential ligands on the protein for complex formation with the heavy-atoms (Sigler & Blow 1965), since the only new derivative in ammonia-free medium, $PdCl_2$, was found to bind in a multiple and non-specific manner (Dickerson, Kopka, Varnum & Weinzierl, 1967*a*).

$PtCl_2$ worked as well as $K_2PtCl_4$, probably because the $PtCl_2$ picked up 2 more ligands to form a square planar complex as soon as it dissolved. The analogous 2 palladous compounds in ammonium sulfate solution did *not* produce intensity changes. Neither did $PdCl_6^{2-}$ or any of the analogous octahedral platinic complexes, with ligands $Cl^-$, $Br^-$, $I^-$, $CN^-$ or $CNS^-$. None of the ammonium, ethylene diammine or diethylene triammine cation complexes with Pd or Pt which were tried worked; this is not surprising, since the protein itself carries a considerable net positive charge in the pH range of stable crystals. The gold complex which is isoelectronic with $PtCl_4^{2-}$, $AuCl_4^-$, did not produce intensity changes, nor did $Au(CN)_2^-$. Iridous and iridic hexahalide complexes failed, as did mercuric tetrachloride and tetraiodide anions. Various other miscellaneous complexes of molybdenum, ruthenium, cadmium, tin, antimony, tungsten, rhenium, osmium, lead, thallium and uranium were tested without success.

One of the problems of such a screening procedure is the pH dependence of binding. Most of the complexes were tried initially either at the pH which resulted naturally upon their addition or at pH 6 to 7 if the natural pH fell outside the stable crystal range. Those which showed promise in our experience or that of other groups were then retested at a series of mole ratios and pH values. Mersalyl was found to bind well to cytochrome *c* at a pH of 6·5 or higher, but poorly below 6·0. Recent work with $OsCl_3$ has shown intensity changes at pH 4·8 but not at 5·3, 5·7 or 6·6 (mole ratios 5 : 1 or 10 : 1). Of the compounds listed above, it cannot be said with assurance that *no* other cytochrome derivative exists, but only that it has not appeared under reasonable testing.

(ii) *Mersalyl analogs and other organomercurials*

After the success of mersalyl and some slight changes in crystals containing parahydroxymercuribenzoate, a series of mersalyl analogs was collected or synthesized and tried with cytochrome *c*. These compounds were chosen to have as common elements a mercury atom, a benzene ring and a negatively charged group, and included parachloromercuribenzoate, parachloromercuribenzene sulfonate, hydroxymercuribenzene, parahydroxymercuriphenoxyacetic acid, 3-hydroxymercuri-4-aminobenzoate, 2-amino-5-hydroxymercuribenzoate, 2-hydroxymercuri-3-nitrobenzoate, 2-amino-5-hydroxymercuribenzene sulfonate and 2-hydroxymercuri-5-aminobenzene sulfonate. None of them produced intensity changes. The work on mersalyl analogs is continuing with heavy-atom modifications of mersalyl itself. Other mercury compounds tried without success include *p*-aminophenylmercuric acetate, phenylmercuric nitrate, mercurochrome, and *p*-chloromercuriphenyl-1-azo-2-naphthol.

(iii) *Carboxymethylation trials*

Following the lead of Okunuki and others (Ando, Matsubara & Okunuki, 1965,1966*a*,*b*), several attempts have been made to label selectively 2 residues, His 33 and Met 65, by carboxymethylation reactions using 2 heavy-atom-bearing bromoacetamide analogs synthesized in this laboratory, *o*-bromomercuribromoacetanilide and 2-bromoacetamido-5-iodobenzoic acid. They have been observed to react with free histidine and methionine, and Okunuki's reaction of the control bromoacetamide with cytochrome has been shown to occur under our experimental conditions. The final link, the reaction of the new labeled compounds with cytochrome, has not yet been successful, and further trials are being made. The reaction is carried out in a solution of reduced cytochrome *c* at pH 5 to 8 under a nitrogen atmosphere at 20 to 37°C over a period of several days; the products are purified on a Bio-Rex 70 column and crystallized. If such an approach can be made to work for proteins in general, it offers several advantages over the traditional diffusion methods:

(a) the heavy-metal would be bound selectively to a known residue and would serve as a marker group in the map;

(b) the group would be covalently bound and would be stable to later diffusion of other metals to form double derivatives; and

(c) being stoichiometrically bound, the group would make the maximum possible phase-determining contribution, and could even be used to place the data on an absolute intensity scale.

To date, the *o*-bromomercuribromoacetanilide preparations have crystallized well but have not shown any changes in X-ray intensities. The 2-bromoacetomido-5-iodobenzoic acid products have not been crystallizable so far. The carboxymethylation experiments will be reported in detail later if they prove to be successful.†

(iv) *Alternative crystal media*

It was suspected that one reason for the failure of most reagents to bind might be competition for metal binding from free $NH_3$ in solution (Sigler & Blow, 1965), and a search was made for ammonia-free media. Lithium chloride, sodium and magnesium sulfate, sodium citrate and lanthanum nitrate were all tried, and rejected because of the impossibility of achieving high enough ionic strength to keep the cytochrome crystals out of solution.

Saturated lanthanum nitrate solution was tried because of its combination of trivalent cation, efficient in building a high ionic strength, and a solubilizing anion. Crystals transferred into lanthanum nitrate from ammonium sulfate were not stable; they swelled and dissolved in a matter of minutes. But in the process they showed a characteristic warping behavior which may permit a tentative guess to be made as to the absolute handedness of the $4_1$ and $4_3$ crystal screw axes. As long crystals (axial ratio about 8 : 1) swelled, they twisted into right-handed helices. When such a long crystal was viewed end-on, its far end was observed to twist in a clockwise direction. This is precisely the swelling and relaxing behavior which would be expected from left-handed helices within the crystal, and suggests that the true space group may be $P4_3$ instead of $P4_1$.

It was found possible to transfer crystals grown in ammonium sulfate to mixed solutions of $Na_2HPO_4/NaH_2PO_4$, with total phosphate molarities of 4·0, 4·3, 4·6 and 5·0 M, a 1 : 1 mole ratio giving a pH of around 5·5 to 6·0. Crystals in 4·0 M-phosphate had to be kept at 5°C to prevent solution; the higher phosphate samples were stable for weeks at

† The Third Derivative

Turning and tumbling in the bubbling stream,
The reactant cannot bear the reagent;
Things fall apart; the crystal cannot hold;
Disorder is loosed upon the world,
The blood-red block is loosed, and everywhere
The symmetry of solid state is drowned;
The best lack of all cohesion, while the worst
Are full of passionate intensity.
Surely some revelation is at hand;
. . . . . .
. . . . . *(with apologies to W. B. Yeats)*

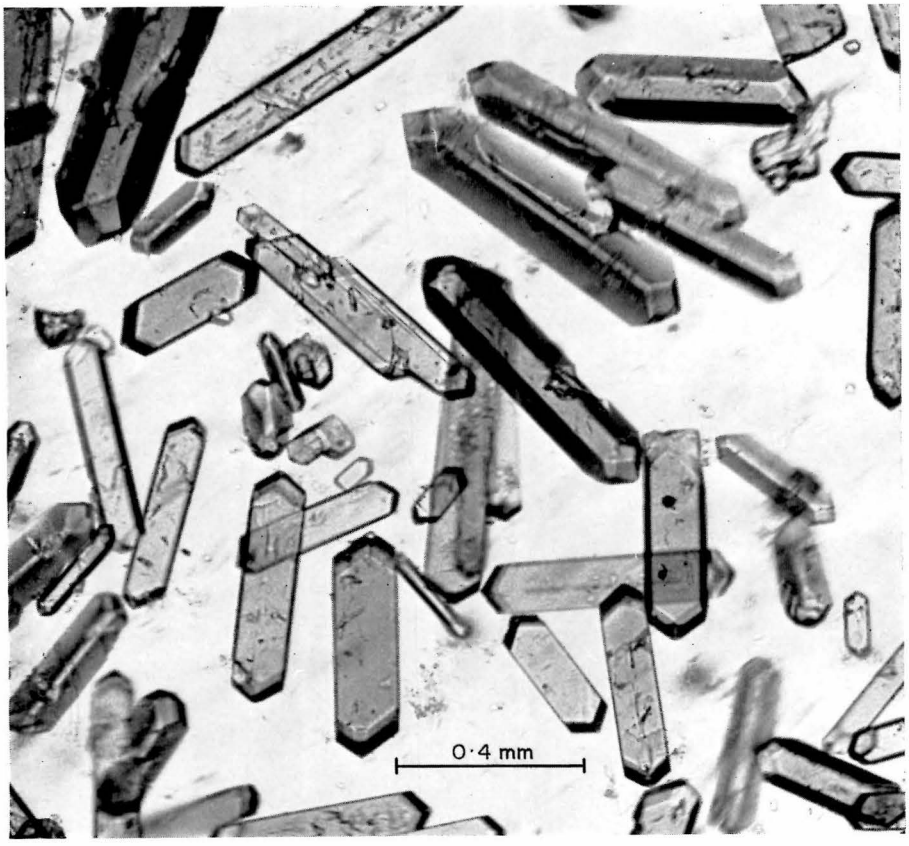
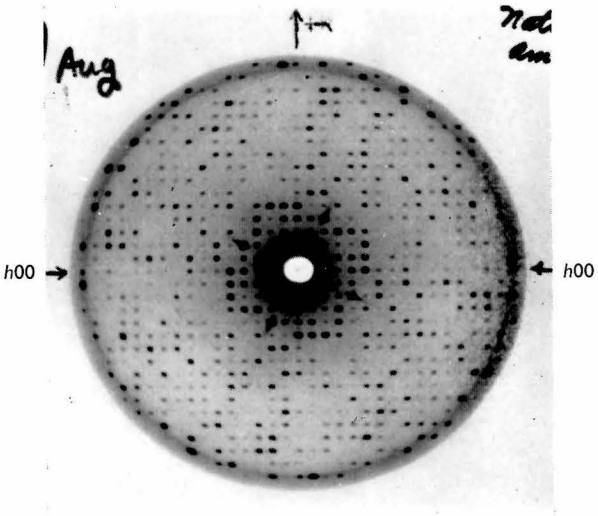
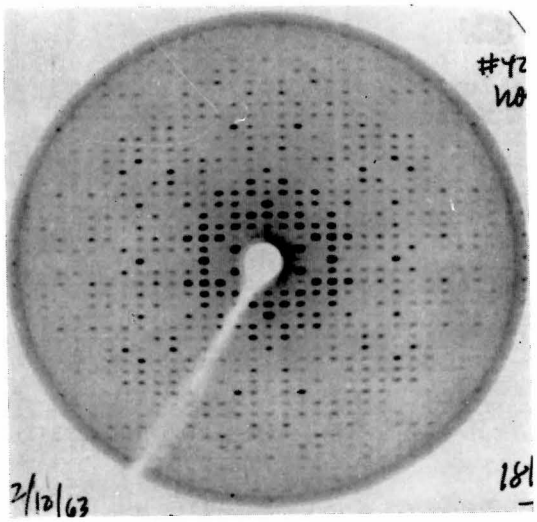



PLATE I. Photomicrograph of crystalline horse heart ferricytochrome *c*. The long axis of the crystals is the *c*-axis, and the *a*- and *b*-axes are normal to the prominent side faces.

[*facing p. 80*

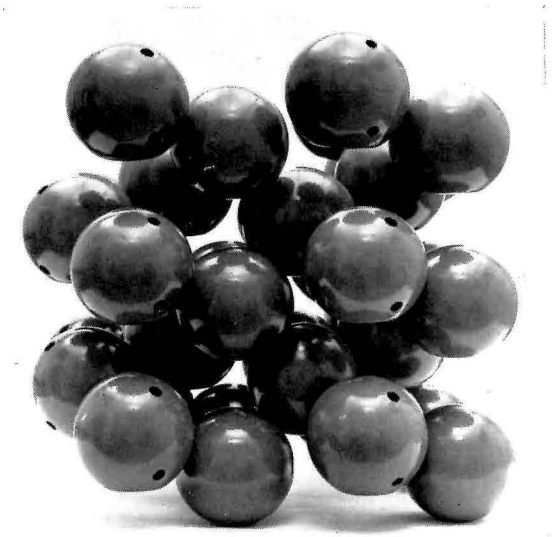



(a)




(b)

PLATE II. Precession camera photographs of native cytochrome *c* crystals.
(a) *hk*0 at 4 Å ($\mu = 12{\cdot}5°$).
(b) *h*0*l*, *l* axis horizontal, at 2·8 Å ($\mu = 17°$).



(a)

(b)

PLATE III. Packed-sphere model of cytochrome *c* crystals based upon a 31 Å sphere and the evidence of Figs 10 and 11.

(a) View down the *x*- or *y*-axis with the *z*-axis vertical.

(b) View down the *z*-axis. The large open channel is at (0, 0) and the small channel is at ($\frac{1}{2}$, $\frac{1}{2}$). The spheres lie on the 120 planes through the crystal.

room temperature. Photographs of crystals in phosphate showed that the X-ray pattern was unchanged except for the inner spots, out to approximately fourth order. These changes were used for the salt-effect studies of section 3(e).

22 different potential heavy-atom derivatives were tried in 4·3 or 5·0 M-phosphate, in 35 photographic trials. Mersalyl and $PtCl_4^{2-}$ gave intensity changes identical to their sulfate changes. Three compounds gave minor and unreproducible changes, but the only new derivative was $PdCl_2$ or $PdCl_4^{2-}$. Both these compounds produced the same extensive changes (Dickerson *et al.*, 1967*a*). But the binding was shown to be either non-specific or of such a nature as to distort the isomorphism of the protein, and the derivative was rejected.

### (d) *Data collection and processing*

The set of platinum data labeled Pt30w was obtained by adding $PtCl_4^{2-}$ in synthetic mother liquor at a final platinum–cytochrome mole ratio of 7·5 : 1 at pH 6, allowing the sample to diffuse for 30 weeks, and then taking an *hk*0 photograph. The set labeled Pt6w is the *hk*0 part of the three-dimensional Pt data. Crystals for these data were prepared at a 1 : 1 mole ratio and aged for 6 weeks before photography. The mersalyl derivative was prepared in a similar manner at a 10 : 1 mole ratio at pH 6·8 and photographed at 1 week (Hg1w) and again at 14 weeks (Hg14w). The double derivative was prepared at pH 6·8 with a Pt : Hg : cytochrome ratio of 1 : 3 : 1 and was photographed at 7 weeks. All data except those for Pt6w were obtained from 12·5° precession camera photographs, using CuK$\alpha$ radiation and a Ni filter. Intensities were measured on a Joyce–Loebl mark III microdensitometer and were corrected for Lorentz polarization factors in the usual way. All computations were carried out on the IBM 7094 computer using locally written programs in FORTRAN IV.

The radial distributions of $F$ data of parent cytochrome, together with changes produced by heavy-atoms, and experimental error are shown in Fig. 1. It can be seen that platinum

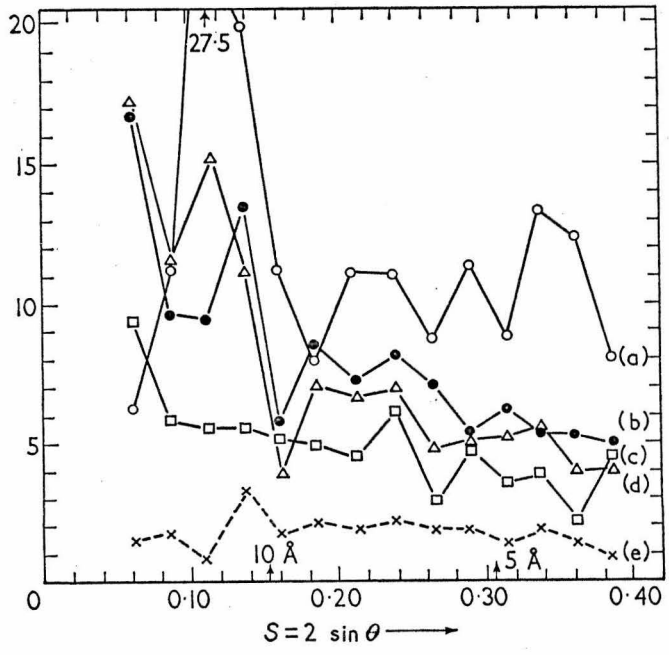



FIG. 1. Radial distribution plots of *hk*0 structure factor data.

(a) Mean native cytochrome structure factor divided by 2. (b) Mean $|\Delta F|$ for Pt30w derivative; (c) mean $|\Delta F|$ for Hg14w; (d) mean $|\Delta F|$ for the double derivative; (e) mean $|\Delta F|$ between two independent native cytochrome *c* films.

is a very strong derivative, the mercury contribution is 2 to 3 times the noise level, and that the double derivative as prepared here is intermediate in quality.

Data were also collected for the parent protein and for Pt and Hg in 4·3 M mixed phosphate buffer and for parent cytochrome in 5·0 M-phosphate. *F* data are to be found in Dickerson *et al.* (1967*a*) for native cytochrome ("N in AS"), Pt30w ("Pt in AS"), Hg14w ("Hg in AS") and the double derivative, all in ammonium sulfate, and for native cytochrome and the Pd derivative in 4·3 M-phosphate. *F* data for Pt6w and Hg1w in sulfate, for Pt and Hg in 4·3 M-phosphate, and for native cytochrome in 5·0 M-phosphate are given in Table 1.

## TABLE 1

*Observed structure factors*

| h k | Pt6 | Hg1 | N5.0 | PtP | HgP | h k | Pt6 | Hg1 | N5.0 | PtP | HgP | h k | Pt6 | Hg1 | N5.0 | PtP | HgP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. 0. | 22.77 | 13.55 | 118.25 | 63.86 | -1.00 | 5. 1. | 65.48 | 56.35 | 67.28 | 63.15 | 60.82 | 9. 6. | 53.53 | 51.32 | 43.51 | 44.58 | 55.70 |
| 1. 1. | 5.33 | 26.19 | 74.91 | 38.60 | -1.00 | 5. 2. | 46.51 | 76.05 | 72.43 | 42.00 | 84.22 | 9. 7. | 4.46 | 3.38 | 5.05 | 4.85 | 5.64 |
| 1. 2. | 2.50 | -1.00 | 29.70 | 17.71 | 26.67 | 5. 3. | 36.43 | 27.36 | 37.53 | 35.52 | 29.68 | 9. 8. | 15.34 | 19.41 | 12.77 | 13.32 | 23.07 |
| 1. 3. | -1.00 | 15.77 | 18.50 | 12.28 | -1.00 | 5. 4. | 13.99 | 4.40 | 7.44 | 16.94 | 4.61 | 9. 9. | 21.85 | 13.21 | 15.95 | 14.08 | 8.90 |
| 1. 4. | -1.00 | 16.64 | 23.87 | 35.82 | 13.53 | 5. 5. | 51.76 | 43.93 | 42.52 | 44.15 | 53.29 | 9.10. | 19.80 | 18.02 | 14.64 | 19.42 | 16.74 |
| 1. 5. | 9.40 | 13.98 | 16.70 | 3.28 | 17.65 | 5. 6. | 46.49 | 28.33 | 28.87 | 36.92 | 24.46 | 9.11. | 5.08 | 5.57 | -1.00 | 4.90 | 7.43 |
| 1. 6. | 40.39 | 40.13 | 51.64 | 36.91 | 40.20 | 5. 7. | 34.73 | 10.41 | 13.76 | 24.67 | 5.25 | 9.12. | 5.77 | 3.02 | -1.00 | 4.72 | 5.16 |
| 1. 7. | 9.78 | 14.16 | 19.58 | 3.92 | 20.97 | 5. 8. | 38.65 | 28.97 | 29.38 | 22.73 | 22.31 | 10. 0. | 7.16 | 3.26 | 6.00 | 5.57 | 5.49 |
| 1. 8. | 60.63 | 55.60 | 54.35 | 45.52 | 45.35 | 5. 9. | 33.12 | 46.76 | 49.51 | 34.99 | 50.85 | 10. 1. | 8.95 | 3.27 | 4.64 | 4.63 | 5.50 |
| 1. 9. | 10.17 | 22.15 | 14.23 | 6.49 | 25.41 | 5.10. | 37.43 | 39.05 | 37.04 | 33.60 | 41.79 | 10. 2. | 26.57 | 4.57 | 4.67 | 26.23 | 5.50 |
| 1.10. | 38.24 | 33.20 | 31.42 | 31.74 | 37.14 | 5.11. | 29.43 | 17.33 | -1.00 | 29.35 | 20.50 | 10. 3. | 3.69 | 3.37 | 4.71 | 4.67 | 5.51 |
| 1.11. | 48.73 | 63.57 | -1.00 | 43.57 | 62.43 | 5.12. | 3.87 | 4.73 | -1.00 | 5.03 | 8.92 | 10. 4. | 24.43 | 32.56 | 40.90 | 25.31 | 27.69 |
| 1.12. | 49.36 | 28.81 | -1.00 | 37.82 | 26.06 | 5.13. | -1.00 | 28.77 | -1.00 | 30.89 | 28.18 | 10. 5. | 18.98 | 14.98 | 12.44 | 21.06 | 17.31 |
| 1.13. | 12.88 | 15.60 | -1.00 | 7.51 | 13.90 | 5.14. | -1.00 | 12.53 | -1.00 | 17.38 | 11.78 | 10. 6. | 23.40 | 26.48 | 26.57 | 17.99 | 23.73 |
| 1.14. | 15.78 | 3.27 | -1.00 | 11.17 | 5.54 | 5.15. | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | 10. 7. | 5.01 | 17.08 | 5.24 | 4.94 | 23.61 |
| 1.15. | 25.49 | 26.53 | -1.00 | 17.27 | 23.48 | 6. 0. | 26.92 | 20.44 | 17.50 | 18.24 | 16.60 | 10. 8. | 46.50 | 50.94 | 49.43 | 38.41 | 50.62 |
| 2. 0. | 29.06 | 20.26. | 19.88 | 16.99 | 35.38 | 6. 1. | 6.11 | 12.16 | 10.88 | 3.63 | 27.70 | 10. 9. | 24.65 | 12.04 | 7.00 | 17.66 | 7.90 |
| 2. 1. | 2.50 | 1.67 | 33.34 | 20.30 | -1.00 | 6. 2. | 2.79 | 5.24 | 3.53 | 3.73 | 4.60 | 10.10. | 30.94 | 28.02 | 29.68 | 26.27 | 30.84 |
| 2. 2. | 11.71 | 9.21 | 47.61 | 15.17 | 21.03 | 6. 3. | 3.84 | 3.39 | 8.97 | 3.82 | 4.71 | 10.11. | 16.96 | 27.38 | -1.00 | 18.15 | 23.87 |
| 2. 3. | 43.64 | 37.85 | 23.46 | 35.06 | 32.69 | 6. 4. | 13.42 | 5.94 | 4.98 | 12.70 | 4.82 | 10.12. | 65.67 | -1.00 | -1.00 | 47.59 | -1.00 |
| 2. 4. | 188.68 | 148.51 | 159.16 | 141.32 | 161.85 | 6. 5. | 24.81 | 28.58 | 25.45 | 19.91 | 27.08 | 11. 0. | 22.14 | 19.46 | -1.00 | 20.02 | 17.91 |
| 2. 5. | 4.75 | 16.61 | 9.88 | 3.37 | 15.87 | 6. 6. | 17.26 | 13.99 | 18.48 | 13.86 | 9.89 | 11. 1. | 30.94 | 4.06 | -1.00 | 28.00 | 5.63 |
| 2. 6. | 50.69 | 47.69 | 45.12 | 44.22 | 40.63 | 6. 7. | 26.23 | 31.34 | 30.48 | 23.24 | 29.28 | 11. 2. | 14.21 | 9.87 | -1.00 | 10.44 | 13.73 |
| 2. 7. | 2.50 | 15.87 | 15.96 | 5.99 | 13.76 | 6. 8. | 17.27 | 12.21 | 8.05 | 15.00 | 10.99 | 11. 3. | 23.25 | 22.28 | -1.00 | 25.65 | 24.41 |
| 2. 8. | 2.51 | 11.08 | 13.85 | 4.24 | 15.43 | 6. 9. | 4.12 | 3.37 | 4.93 | 7.75 | 5.62 | 11. 4. | -1.00 | 3.39 | -1.00 | 4.88 | 5.65 |
| 2. 9. | 45.43 | 39.90. | 47.17 | 42.58 | 34.04 | 6.10. | 3.83 | 4.08 | 6.57 | 4.88 | 8.94 | 11. 5. | 26.23 | 32.05 | -1.00 | 23.18 | 35.64 |
| 2.10. | 32.48 | 15.32 | 21.02 | 32.69 | 16.70 | 6.11. | 32.51 | 39.48 | -1.00 | 32.55 | 40.71 | 11. 6. | 58.07 | 61.14 | -1.00 | 53.71 | 55.80 |
| 2.11. | 21.63 | 21.12 | -1.00 | 18.51 | 26.63 | 6.12. | 45.51 | 51.07 | -1.00 | 47.49 | 52.09 | 11. 7. | -1.00 | 28.40 | -1.00 | 39.45 | 23.23 |
| 2.12. | 9.39 | 21.69 | -1.00 | 6.63 | 23.71 | 6.13. | 2.81 | 9.47 | -1.00 | 4.92 | 12.14 | 11. 8. | 12.83 | 3.35 | -1.00 | 5.01 | 5.62 |
| 2.13. | 69.05 | 52.37 | -1.00 | 58.65 | 50.54 | 6.14. | -1.00 | 38.44 | -1.00 | 32.60 | 37.65 | 11. 9. | 25.97 | 23.08 | -1.00 | 22.07 | 24.16 |
| 2.14. | 72.80 | 60.46 | -1.00 | 60.16 | 61.68 | 7. 0. | 39.59 | 16.61 | 16.59 | 34.39 | 11.04 | 11.10. | 9.57 | 3.11 | -1.00 | 4.71 | 5.15 |
| 2.15. | 31.06 | 24.42 | -1.00 | 25.15 | 24.66 | 7. 1. | 34.30 | 17.55 | 24.92 | 35.56 | 18.67 | 11.11. | -1.00 | -1.00 | -1.00 | 25.95 | -1.00 |
| 3. 0. | 9.77 | 22.97 | 3.78 | 4.88 | 22.13 | 7. 2. | -1.00 | 10.86 | 6.23 | 26.81 | 14.26 | 12. 0. | 27.17 | 7.92 | -1.00 | 27.51 | 11.94 |
| 3. 1. | -1.00 | 35.47 | 15.28 | 17.01 | 26.70 | 7. 3. | 2.71 | 3.57 | 3.91 | 4.04 | 4.93 | 12. 1. | 4.48 | 3.39 | -1.00 | 5.54 | 5.66 |
| 3. 2. | 40.13 | 17.12 | 14.90 | 37.03 | 23.25 | 7. 4. | 37.15 | 21.88 | 27.17 | 41.04 | 24.29 | 12. 2. | 41.92 | 22.66 | -1.00 | 36.40 | 25.05 |
| 3. 3. | 74.82 | 64.38 | 59.84 | 63.07 | 63.37 | 7. 5. | 3.66 | 11.01 | 15.83 | 13.21 | 12.02 | 12. 3. | 27.34 | 24.52 | -1.00 | 27.20 | 22.29 |
| 3. 4. | 34.48 | 15.66 | 4.61 | 28.67 | 8.17 | 7. 6. | 8.61 | 3.17 | 4.38 | 6.34 | 5.37 | 12. 4. | 14.87 | 29.64 | -1.00 | 10.25 | 29.68 |
| 3. 5. | 18.53 | 32.41 | 24.41 | 12.24 | 29.48 | 7. 7. | 4.94 | 3.27 | 4.62 | 4.62 | 5.49 | 12. 5. | 31.86 | 35.49 | -1.00 | 28.71 | 33.07 |
| 3. 6. | 19.93 | 14.81 | 17.81 | 29.05 | 20.14 | 7. 8. | 39.64 | 39.00 | 31.45 | 31.49 | 38.69 | 12. 6. | 20.76 | 32.09 | -1.00 | 19.86 | 32.86 |
| 3. 7. | 25.94 | 36.96 | 39.15 | 24.96 | 28.16 | 7. 9. | 21.02 | 18.03 | 22.08 | 20.36 | 17.33 | 12. 7. | -1.00 | 3.35 | -1.00 | 4.96 | 5.53 |
| 3. 8. | 9.10 | 3.08 | 4.27 | 11.11 | 5.15 | 7.10. | 32.34 | 23.35 | 21.10 | 23.06 | 22.68 | 12. 8. | -1.00 | 28.26 | -1.00 | 31.21 | 26.30 |
| 3. 9. | -1.00 | 3.94 | 15.79 | 29.46 | 11.16 | 7.11. | -1.00 | 10.44 | -1.00 | 18.95 | 8.92 | 12. 9. | 51.94 | 33.38 | -1.00 | 33.83 | 34.10 |
| 3.10. | 3.69 | 4.05 | 4.71 | 4.67 | 9.48 | 7.12. | 47.83 | 32.47 | -1.00 | 32.83 | 34.54 | 12.10. | -1.00 | -1.00 | -1.00 | 30.71 | -1.00 |
| 3.11. | 4.25 | 3.38 | -1.00 | 9.70 | 5.64 | 7.13. | -1.00 | 15.19 | -1.00 | 14.94 | 19.86 | 13. 0. | 12.55 | 8.33 | -1.00 | 14.90 | 7.98 |
| 3.12. | 36.64 | 37.51 | -1.00 | 35.31 | 30.45 | 7.14. | -1.00 | -1.00 | -1.00 | 4.32 | -1.00 | 13. 1. | 4.73 | 5.43 | -1.00 | 5.03 | 5.70 |
| 3.13. | 29.04 | 33.65 | -1.00 | 25.36 | 31.59 | 8. 0. | 27.64 | 40.69 | 33.23 | 24.27 | 43.92 | 13. 2. | 31.14 | 29.47 | -1.00 | 28.09 | 24.08 |
| 3.14. | 53.62 | 67.09 | -1.00 | 51.17 | 70.56. | 8. 1. | 22.12 | 13.97 | 17.73 | 27.61 | 16.12 | 13. 3. | 14.63 | 18.10 | -1.00 | 10.88 | 17.87 |
| 3.15. | 6.23 | 16.68 | -1.00 | 10.11 | 21.53 | 8. 2. | 5.64 | 12.78 | 22.71 | 10.48 | 9.58 | 13. 4. | 44.39 | 23.86 | -1.00 | 36.85 | 23.95 |
| 4. 0. | 56.34 | 47.96 | 36.50 | 49.33 | 46.93 | 8. 3. | 13.07 | 7.17 | 5.44 | 16.12 | 5.86 | 13. 5. | -1.00 | 71.56 | -1.00 | 56.45 | 71.60 |
| 4. 1. | -1.00 | 25.71 | 42.27 | 30.43 | 22.78 | 8. 4. | 19.07 | 32.06 | 20.22 | 10.09 | 29.55 | 13. 6. | -1.00 | 19.91 | -1.00 | 17.97 | 13.50 |
| 4. 2. | 25.57 | 16.99 | 2.93 | 21.22 | 3.88 | 8. 5. | 32.97 | 22.03 | 28.70 | 33.67 | 21.10 | 13. 7. | -1.00 | 4.35 | -1.00 | 4.79 | 5.25 |
| 4. 3. | 55.23 | 51.70 | 50.36 | 41.68 | 55.12 | 8. 6. | 4.00 | 7.65 | 12.68 | 4.63 | 10.28 | 13. 8. | -1.00 | -1.00 | -1.00 | 4.56 | 4.96 |
| 4. 4. | 17.89 | 7.17 | 13.06 | 14.15 | 4.39 | 8. 7. | 70.18 | 66.10 | 58.83 | 59.51 | 71.26 | 14. 0. | 36.52 | 24.46 | -1.00 | 38.92 | 23.91 |
| 4. 5. | 4.87 | 9.98 | 7.93 | 9.48 | 10.91 | 8. 8. | 23.45 | 4.07 | 16.39 | 21.99 | 5.64 | 14. 1. | 25.19 | 17.40 | -1.00 | 23.19 | 18.28 |
| 4. 6. | 6.65 | 12.56 | 14.93 | 12.94 | 27.01 | 8. 9. | 29.64 | 17.76 | 17.82 | 24.05 | 15.96 | 14. 2. | 22.40 | 18.21 | -1.00 | 15.91 | 17.56 |
| 4. 7. | 23.80 | 22.31 | 31.00 | 26.45 | 23.18 | 8.10. | 43.49 | 32.47 | 31.57 | 32.40 | 34.06 | 14. 3. | 5.74 | 11.59 | -1.00 | 4.92 | 14.66 |
| 4. 8. | 33.42 | 33.18 | 34.65 | 25.73 | 26.81 | 8.11. | 28.13 | 15.84 | -1.00 | 16.52 | 12.18 | 14. 4. | 13.67 | 3.19 | -1.00 | 12.47 | 5.35 |
| 4. 9. | 28.52 | 27.58 | 37.02 | 34.35 | 26.33 | 8.12. | -1.00 | 5.14 | -1.00 | 16.32 | 5.34 | 14. 5. | -1.00 | 3.11 | -1.00 | 4.71 | 5.15 |
| 4.10. | 8.36 | 6.30 | 4.92 | 6.52 | 5.62 | 8.13. | -1.00 | -1.00 | -1.00 | 51.47 | 59.50 | 14. 6. | -1.00 | 20.24 | -1.00 | 17.27 | 13.82 |
| 4.11. | -1.00 | 3.39 | -1.00 | 14.55 | 5.65 | 9. 0. | 6.82 | 3.17 | 4.33 | 14.06 | 5.26. | 14. 7. | -1.00 | -1.00 | -1.00 | 15.62 | -1.00 |
| 4.12. | 10.60 | 3.41 | -1.00 | 4.99 | 5.69 | 9. 1. | 34.39 | 31.13 | 23.85 | 35.30 | 38.98 | 15. 0. | 22.43 | 8.64 | -1.00 | 24.58 | 9.79 |
| 4.13. | 17.46 | 18.48 | -1.00 | 19.12 | 18.95 | 9. 2. | 27.79 | 35.10 | 40.41 | 27.04 | 36.58 | 15. 1. | 5.22 | 9.61 | -1.00 | 6.95 | 14.53 |
| 4.14. | 18.45 | 22.80 | -1.00 | 13.33 | 24.96 | 9. 3. | -1.00 | 25.81 | 19.53 | 12.05 | 36.52 | 15. 2. | 24.30 | 3.03 | -1.00 | 19.83 | 5.06 |
| 4.15. | 18.48 | -1.00 | -1.00 | 15.66 | -1.00 | 9. 4. | 24.00 | 11.34 | 13.96. | 23.33 | 10.88 | 15. 3. | 17.02 | -1.00 | -1.00 | 19.54 | 14.03 |
| 5. 0. | 60.61 | 65.26 | 76.37 | 48.95 | 57.11 | 9. 5. | 32.07 | 33.75 | 36.96 | 26.97 | 31.79 | 15. 4. | 41.22 | -1.00 | -1.00 | 32.58 | -1.00 |

Observed structure factors for Pt6w, Hg1w, parent cytochrome in 5·0 M-phosphate buffer (N5.0), Pt and Hg in 4·3 M-phosphate buffer (PtP and HgP). Data are to the same scale as Table 1 of Dickerson *et al.* (1967*a*). A value of −1·00 indicates that the reflection intensity was not measured.

## 3. Results and Structure Analysis

### (a) *Heavy-atom location and sign determination*

The different sets of data were put on the same relative scale by applying a scale factor of the form: $K \exp(-DS^2)$ to the derivative data, where $S = 2 \sin \theta$ and the scale constants $K$ and $D$ for each derivative were obtained from Wilson-type plots plots of $\ln \{\langle F_P \rangle / \langle F_{PH} \rangle\}$ *versus* $S^2$. ($\boldsymbol{F}_P$ will be used for the native cytochrome structure factor, $\boldsymbol{F}_{PH}$ for the heavy-atom derivative and $\boldsymbol{f}_H$ for the contribution of the heavy-atom group alone. Bold-face type will indicate the vector quantities with phases or signs; normal type will indicate magnitudes only.) This so-called "level scaling" was then modified by an additional scale constant, $K' = \sum_{h,k} F_P^2 / \sum_{h,k} F_P F_{PH}$, applied to the heavy-atom data to produce what is called "Kraut scaling" and to allow for the increased intrinsic scattering power of the protein with added heavy-atoms (Kraut, Sieker, High & Freer, 1962).

Heavy-atom positions for mercury and platinum were found from

$$(\Delta F)^2 = (|F_{PH}| - |F_P|)^2$$

difference Patterson maps (Fig. 2). The validity of the two individual derivatives was confirmed by the successful interpretation of the double-derivative Patterson map, Fig. 2(c), including Pt–Hg cross vectors. Heavy-atom structure factors were calculated from:

$$\boldsymbol{f}_H = \sum_H 2A \exp(-BS)\{\cos 2\pi(hx + ky) + \cos 2\pi(hy - kx)\}. \quad (1)$$

The effective atomic number, $A$, and the radial falloff factor, $B$, were found by first assuming values of 1·0 and 0·0, and then making a Wilson-type plot of

$$\ln\{\langle|\Delta F|\rangle / \langle|f_H|\rangle\} \textit{ versus } S.$$

The use of $S$ rather than $S^2$ implies that the density profile in real space is broader and flatter than a Gaussian profile. This is reasonable at 4 Å in view of the compound nature of the heavy group, and was found to give a better straight line than a plot against $S^2$. Initial unrefined heavy-atom parameters are given in Table 2(a). The mean discrepancies between observed $\Delta F$ and calculated structure factors,

$$\langle \epsilon \rangle = \langle ||\Delta F| - |f_c|| \rangle$$

were plotted against $S = 2 \sin \theta$ as a measure of isomorphism and the goodness of fit of the heavy-atom models. The curve for Hg fell almost exactly atop the experimental error curve, Fig. 3(e), and that for $PtCl_4^{2-}$ was about twice the height of the error curve, indicating that the single-site model fits Hg better than Pt. Attempts to find the chloride atoms in two or three dimensions at 4 Å have been inconclusive, but more exact models of the heavy-atom groups will clearly be necessary at higher resolution.

A set of protein signs was obtained separately from each derivative, first by inspection and then using the centroid phase method (Blow & Crick, 1959; Dickerson, Kendrew & Strandberg, 1961*b*). Each set of signs was then used with values of $\Delta F$ from the *other* derivative to calculate a heavy-atom $\Delta F$ difference Fourier map. Such a cross-sign difference map provides a stringent derivative check which a map using signs influenced by the derivative in question cannot do (Dickerson *et al.*, 1967*a*). Figure 3(a) to (c) shows the platinum map with mercury signs, the mercury map with platinum signs, and the double derivative map with the best refined set of signs

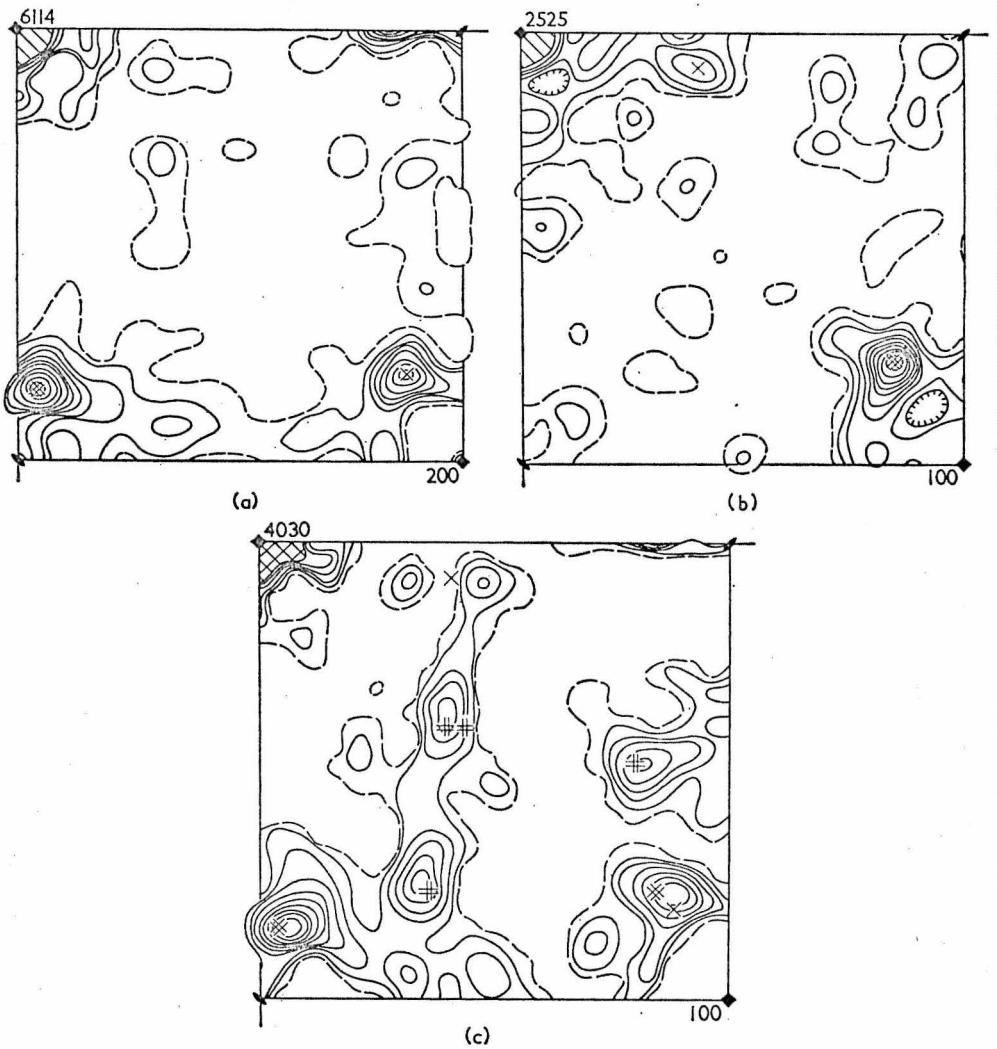



FIG. 2. $hk0$ $(\Delta F)^2$ difference Patterson projections for heavy-atom derivatives at 4 Å resolution.

All maps are to the same arbitrary scale. Contour intervals are marked in the lower right corner and the height of the origin peak is given at the upper left. The zero contour is dashed. Co-ordinates $u$ (horizontal) and $v$ (vertical) run from 0 to $\frac{1}{2}$, with the origin in the upper left corner. (a) $PtCl_4^{2-}$ at 30 weeks; (b) mersalyl at 14 weeks; (c) double derivative, Pt and Hg, at 7 weeks. Single and double Patterson peaks expected from the sites of Table 2(a) are shown by single and double X's. Pt–Hg cross-vectors in Fig. 1(c) are marked by double crosses.

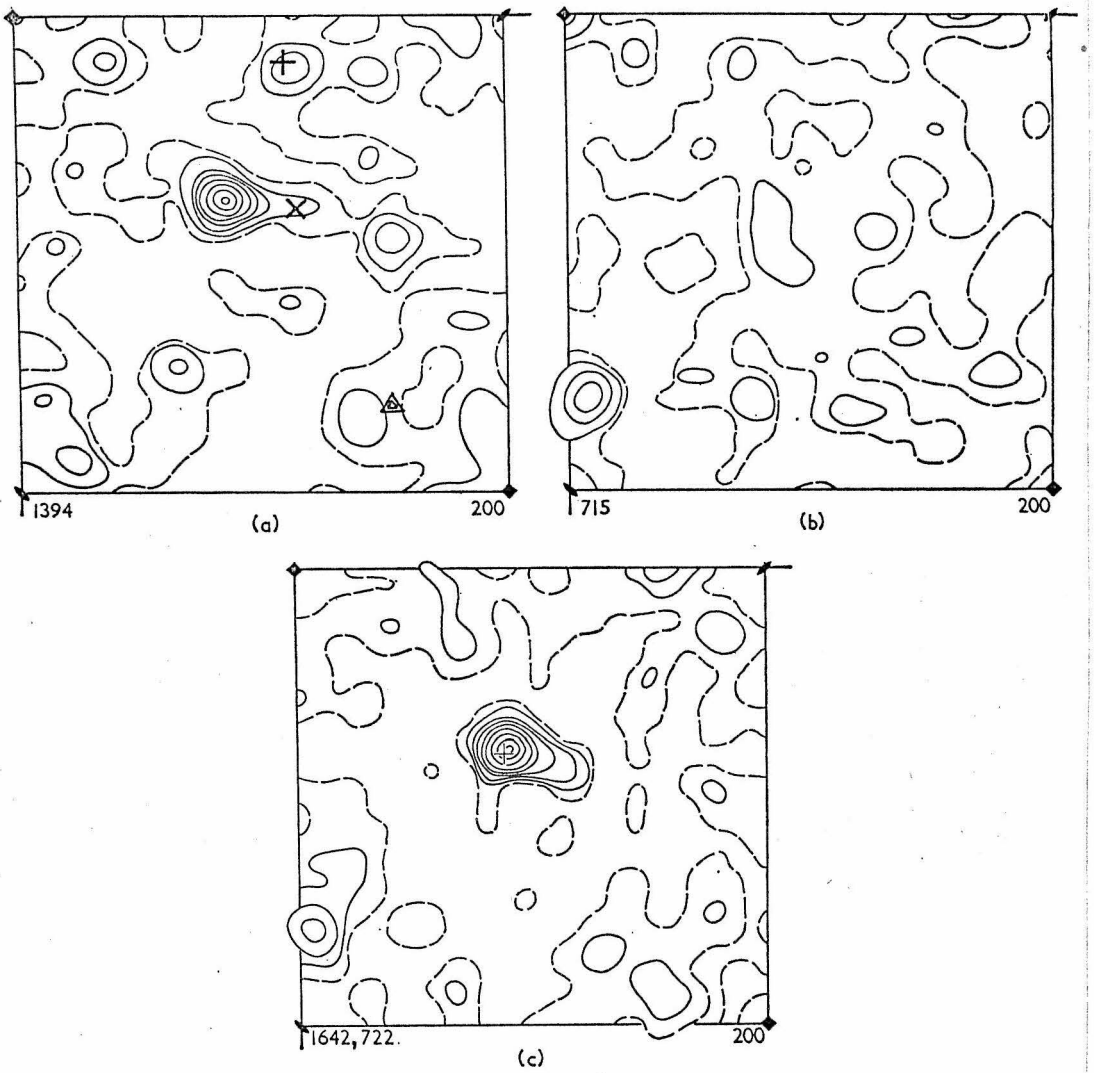



FIG. 3. Cross-difference Fourier maps using $\Delta F$ values from one derivative and protein signs from a *different* derivative, as a check on derivative validity.

All three maps are on the same arbitrary scale, with contour interval at the lower right and height of the principal peak or peaks at the lower left. Zero contours are dashed, negative contours omitted. Co-ordinates $x$ (horizontal) and $y$ (vertical) run from 0 to $\frac{1}{2}$, with the origin in the upper left corner. The maps are not weighted with figures of merit.

(a) Pt $\Delta F$'s and Hg protein signs, showing the true Pt site. The true (X), questionable (+) and false ($\Delta$) secondary sites are marked. (b) Hg $\Delta F$'s and Pt protein signs, showing the true Hg site. (c) Double derivative $\Delta F$'s and four-data protein signs, showing the Pt site (+) and the much weaker Hg site.

TABLE 2

| | $x$ | $y$ | $A$ | $B$ | $K$(Kraut) | $E_j$† |
|---|---|---|---|---|---|---|
| (a) *Initial heavy-atom parameters from difference Patterson maps and Wilson plots* | | | | | | |
| Pt6w | 0·216 | 0·196 | 7·00 | 1·66 | 0·91 | 6·40 |
| Pt30w | 0·220 | 0·200 | 8·15 | 2·42 | 1·12 | 4·90 |
| Hg1w | 0·012 | 0·402 | 5·00 | 2·50 | 1·00 | 3·28 |
| Hg14w | 0·020 | 0·400 | 5·80 | 2·80 | 1·04 | 3·21 |
| Double-Pt | 0·220 | 0·200 | 7·40 | 2·83 | (1·00) | 5·55 |
| Double-Hg | 0·020 | 0·400 | 3·22 | 2·83 | (1·00) | 5·55 |
| (b) *Final heavy-atom parameters from refinement run S37* | | | | | | |
| Pt6w I° | 0·219 | 0·199 | 6·51 | 1·66 | 0·94 | 4·58 |
| II° | 0·271 | 0·204 | 2·79 | 1·66 | 0·94 | |
| Pt30w I° | 0·218 | 0·200 | 7·99 | 2·42 | 1·11 | 4·15 |
| II° | 0·268 | 0·204 | 3·50 | 2·42 | 1·11 | |
| Hg1w | 0·020 | 0·399 | 4·04 | 2·50 | 1·02 | 3·42 |
| Hg14w | 0·021 | 0·400 | 5·33 | 2·80 | 1·06 | 3·03 |
| Double: | | | | | | |
| Pt I° | 0·216 | 0·196 | 6·49 | 2·83 | 1·05 | 5·06 |
| Pt II° | 0·269 | 0·205 | 3·95 | 2·83 | 1·05 | |
| Hg | 0·015 | 0·398 | 2·39 | 2·83 | 1·05 | |

† Root-mean-square error estimates taken from preliminary structure factor calculations.

obtained from the simultaneous use of Pt6w, Pt30w, Hg1w and Hg14w (called the "four-data" set). The appearance of such cross-sign maps is poorer than that of self-sign maps, but the latter are useless for verification purposes.

(b) *Improvement of sign analysis with multiple derivatives*

In principle, in a centric projection only one derivative is needed. As several derivatives are added together, however, the sign analysis will improve markedly because of the averaging down of the effect of data errors and errors in interpreting the individual derivatives. In order to see the significance of this co-operative improvement, signs were first determined individually using the centroid phase program and then with increasingly larger combinations of Pt6w, Pt30w, Hg1w, Hg14w and the double derivative. The effect upon mean figure of merit is shown in Fig. 4†, and upon Kraut $R$ factor in Fig. 5.

The two platinum derivatives are seen to be comparable, and the two mercury derivatives are themselves comparable and somewhat better than the platinum. This is so in spite of the greater intrinsic contribution of Pt, and reflects both the

† The mean figure of merit is a treacherous measure of derivative quality because of its strong dependence upon the choice of the root-mean-square error, $E_j$, for each derivative. An underestimation of the $E_j$'s leads to a spuriously high value of $\langle m \rangle$. The question of weights, $E_j$'s and refinement criteria is considered in Dickerson, Weinzierl & Palmer (1967*b*). A reasonable choice of $E_j$ for each derivative, $j$, in a cycle of phase refinement is:

$$E_j^2 = \langle \epsilon_j^2(\phi_B) \rangle_h$$

where $\epsilon_j(\phi_B)$ is the lack of closure error for that derivative at the centroid phase angle from the previous phase cycle, and the average is taken over all reflections. If the $E_j$'s are adjusted in this manner, then the mean figures of merit are at least on the same relative scale and are comparable with one another. All figures of merit quoted in this paper are after such an adjustment of $E_j$'s.

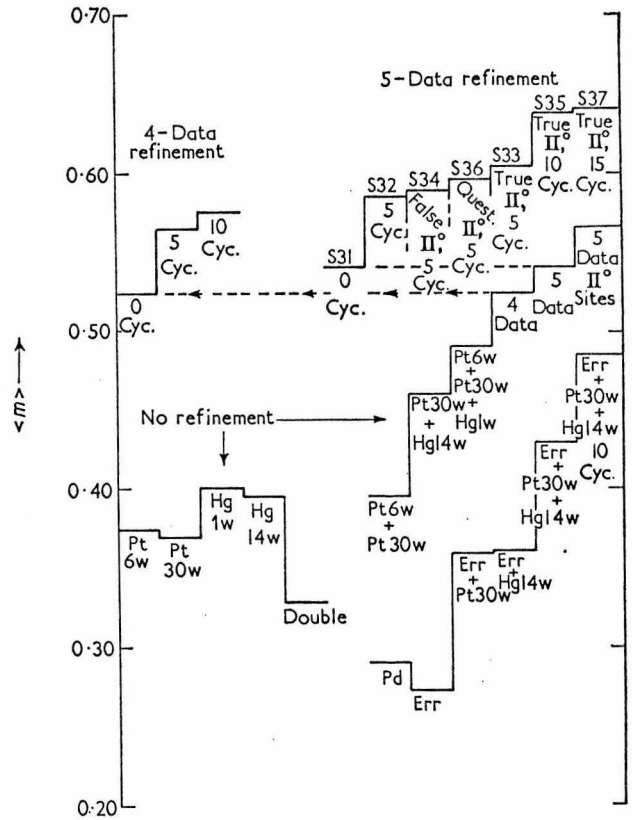



FIG. 4. Mean figures of merit for centroid phase (sign) analyses of the 190 reflections at 4 Å.

Individual derivatives and various combinations are shown, without and with phase refinement. "4-Data" means Pt6w, Pt30w, Hg1w and Hg14w; "5-Data" means these plus the double derivative.

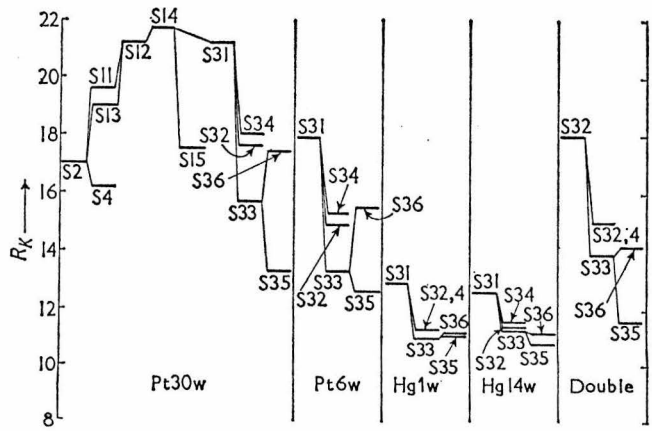



FIG. 5. Kraut $R_K$ factors for individual derivatives.

Run S2 is with Pt alone, S4 is the same run after five refinement cycles. Run S11 is with Pt6w and Pt30w, S13 is with Pt30w and Hg14w, and S12 is with Pt6w, Pt30w and Hg1w, all unrefined. S14 is the four-data set before refinement and S15 is the same set after ten refinement cycles. S31 is the unrefined five-data set and S32 is the same set after five cycles. S33 is the result of five cycles of refinement in the presence of the true secondary Pt site and S34 is the same but with the false secondary site instead. S35 is a continuation of the true site S33 refinement for five more cycles and S36 is a similar continuation of S33 but with the true secondary site now replaced with the questionable one (see text and Fig. 3(a)).

better quality of the Hg data and the closer isomorphism of the Hg derivative. The double derivative is the poorest of all. The derivative which is known to be incorrect (Pd) and an entirely erroneous derivative (Err) obtained by coupling the Pd data with a single site selected at random (Dickerson *et al.*, 1967*a*), both give mean figures of merit significantly worse than even the worst valid derivative.

In general, the addition of another version of the same derivative raises the mean figure of merit by 2·5 to 3·0 percentage points (which merely reflects a reinforcement of the initial derivative, whether right or wrong), the addition of the double derivative to an analysis already containing both sites (which in a sense is much the same thing) raises it by 2%, and the addition of a new derivative at a different site raises $\langle m \rangle$ by 6·5 to 9·5 percentage points.

### (c) *Secondary sites and refinement*

Refinement of heavy-atom parameters $x$, $y$, $A$, $B$, and the derivative data scale factors, $K$, was carried out using a method of minimization of the weighted sum of lack of closure errors which was first used for myoglobin (Dickerson *et al.*, 1961*a*,*b*), and which has since been used successfully with chymotrypsinogen (Kraut *et al.*, 1962), hemoglobin (Muirhead, personal communication) and carboxypeptidase (Lipscomb *et al.*, 1966). Details of the method and program are to be found in the paper by Dickerson *et al.* (1967*b*). Each cycle of operation involves a centroid phase analysis followed by one cycle of full-matrix least squares refinement of the heavy-atom parameters holding phases fixed at their just-determined values. Centroid or "best" phases and not the most probable phases are used. The program, written for the general three-dimensional case, treats centric reflections in the same way as acentric ones, a simpler although less elegant method than the use of a tanh formula (Blow & Crick, 1959). The net result in a centric zone is to produce the same signs but a systematically lower figure of merit, a relatively unimportant effect in the long run.

Ten cycles of refinement of the four-data set took the mean figure of merit from 0·52 to 0·58 and five cycles took the five-data set (with the double derivative added) from 0·54 to 0·59, both from the initial parameters of Table 2(a). Refinement was virtually at an end at the end of each run. It was not possible to refine the radial falloff constant, $B$, at such a low resolution, as $A$ and $B$ interacted too strongly and blew up together. If $B$ was held fixed at its best Wilson-plot value, then $A$ refined normally.

All of the Pt maps, no matter with what sign set, showed a characteristic elongation of the Pt peak to the right as in Fig. 3(a). In some but not all maps, there was a minor peak at + in Fig. 5(a), and this was called the "questionable" secondary site. Finally, a completely false secondary site was chosen at Δ in Fig. 3(a). A comparison was then made of the relative effects of true, questionable and false secondary sites on two refinement criteria, mean figure of merit and Kraut $R$ factor (Kraut *et al.*, 1962):

$$R_K = \sum_{hkl} |\epsilon_{hkl}| / \sum_{hkl} |F_{PH}| \tag{2}$$

Here $\epsilon_{hkl}$ is the lack of closure error of the phase triangle of the derivative in question, using the protein phase angle determined by all the derivatives together:

$$|\epsilon_{hkl}| = ||\boldsymbol{F}_{PH}| - |\boldsymbol{F}_P + \boldsymbol{f}_H|| \tag{3}$$

The results of these comparisons are shown in Fig. 4 for mean figure of merit and in

Fig. 5 for Kraut $R$ factor. Particular attention should be paid to the difference in behavior of $\langle m \rangle$ and $R_K$ upon the addition of the false sites. The mean figure of merit is improved slightly no matter whether the new site is good or bad, simply because there are more parameters to adjust and more degrees of freedom for refinement. $R_K$, in contrast, differentiates clearly between true and false secondary sites. In Fig. 5, for example, run S34 is clearly worse than S32, whereas S33 is better; and again, S36 is clearly inferior to both S33 and S35.

Final heavy-atom parameters, from run S37, are given in Table 2(b). Refinement by this time had come to a complete halt.

TABLE 3

*Kraut* $R_K$ *factors for derivatives*

| | Pt6w | Pt30w | Hg1w | Hg14w | Double |
|---|---|---|---|---|---|
| Initial unrefined Table 2(a) values, individual phase determinations | 15·8% | 17·0 | 10·9 | 10·6 | 17·0 |
| Final 5-data refined Table 2(b) values | 12·6% | 13·2 | 11·1 | 10·7 | 13·8 |

(d) *The* hk*0 Fourier projection*

The *hk*0 Fourier projection resulting from refinement run S37 is shown in Fig. 6. The map is weighted with figures of merit, but the unit-weight map is almost identical except for a scale change. One molecule in projection appears as a ring of high density around a negative region, approximately centered on the map as shown, and with a less deep negative region around it. The 4-fold screw axis projected at $(x, y) = (\frac{1}{2}, \frac{1}{2})$ appears quite negative, while the region around (0, 0) for some distance is relatively flat and featureless. The Pt site appears at the upper edge of the molecule, while the Hg site appears to the left in what would be a region of overlap of the two molecules related across the screw dyad at $(0, \frac{1}{2})$.

In 95% saturated ammonium sulfate, the electron density of the medium, 0·407 e/Å$^3$, is almost exactly equal to the mean electron density to be expected through the protein molecule, judging from the myoglobin and hemoglobin results (Cullis, Muirhead, Perutz & Rossmann, 1962). If the (000) term is omitted from the Fourier synthesis, then this *average* density will be at the zero contour level. Regions of loose van der Waals packing of side chains will show up in three dimensions as negative regions, perhaps approaching the figure of crystalline benzene of 0·29 e/Å$^3$. Regions of tight covalently bound polypeptide framework will appear positive at 4 Å, perhaps up to the upper limit observed for α-helix in 5·5 Å hemoglobin of 0·9 e/Å$^3$.

A model structure for the cytochrome *c* molecule can be developed which explains

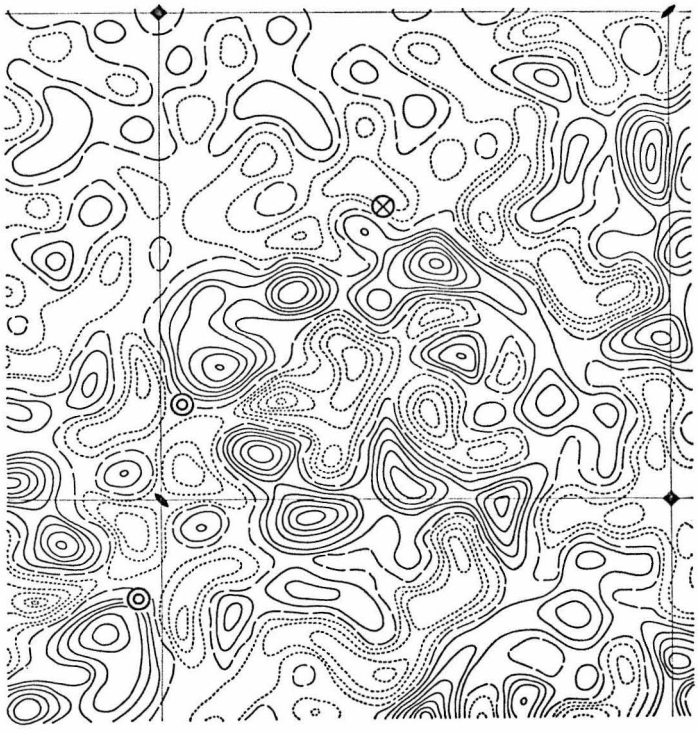


FIG. 6. The $hk0$ projection at 4 Å resolution of horse heart ferricytochrome c, using signs from the best refined parameters of Table 2(b).

The map is weighted with figures of merit. Axis conventions are as in Fig. 3. The contours are at intervals of 300 on the scale of Fig. 3, with the zero contour (dashed) marking the *average* projected electron density throughout the cell. Positive contours are solid and negative contours are dotted. ⊗ marks the Pt site and ◎ the Hg site.

the gross features of Fig. 6 well. In this model, the center of the molecule consists of loosely packed hydrophobic side chains, of negative map density. Around this is a framework of some type, composed of covalently bonded polypeptide chain and forming a shell of positive map density. On the outside of this shell is another negative layer of loosely packed side chain, and surrounding the entire molecule is the salt medium, of average electron density and therefore at or near the zero contour.

The postulated negative center and positive shell show up well in projection in Fig. 6. The very negative regions between two molecules related by the 4-fold screw axis arise from projection down through the external side-chain regions of both molecules, and the negative region around the screw dyad arises in a similar way. The featureless region around the origin is a projection through salt medium alone, and the negative region at $(\frac{1}{2}, \frac{1}{2})$ indicates an overlap in projection of side chains of the molecules related by the screw tetrad. Additional support for this model is provided by a study of the effect of salt medium density on the diffraction pattern.

### (e) *Salt effect study*

Upon examining the diffraction patterns of crystals in 95% saturated ammonium sulfate, 4·3 M and 5·0 M-mixed phosphate buffer, it was found that the innermost

intensities out to about the fourth order were altered, but that beyond this region the intensity changes were within the experimental uncertainties. This is to be expected if the molecular structure and orientation are unchanged and if the only difference is in the scattering density of the medium between molecules. The only reflections affected are then those low-order terms needed to outline the molecules. This fact was used to prepare a salt difference Fourier map, using as coefficients the quantities: $(F_{4\cdot3P} - F_{AS})$, where the two structure factors are those of the parent cytochrome in 4·3 M-phosphate and in ammonium sulfate, *with their correct sign* in each medium.

The signs in 4·3 M-phosphate were found from plots of $F$ *versus* electron density of the medium. The structure factor itself must be a linear function of the mean electron density of the crystallizing medium. For if the cell volume is divided into two regions, molecule and surroundings, then the total scattering factor is the sum of the transform of the molecular density integrated over the molecular volume and the transform of the density of the uniform surrounding medium integrated over the non-molecule volume:

$$F_{(S)} = \int_{\text{molec.}} \rho_{(r)} \exp 2\pi i\, \mathbf{S} \cdot \mathbf{r}\, d_{\mathbf{r}} + \rho_{\text{salt}} \int_{\text{medium}} \exp 2\pi i\, \mathbf{S} \cdot \mathbf{r}\, d_{\mathbf{r}} \qquad (4)$$

If the constant salt density is removed from the second integral as in equation (4)

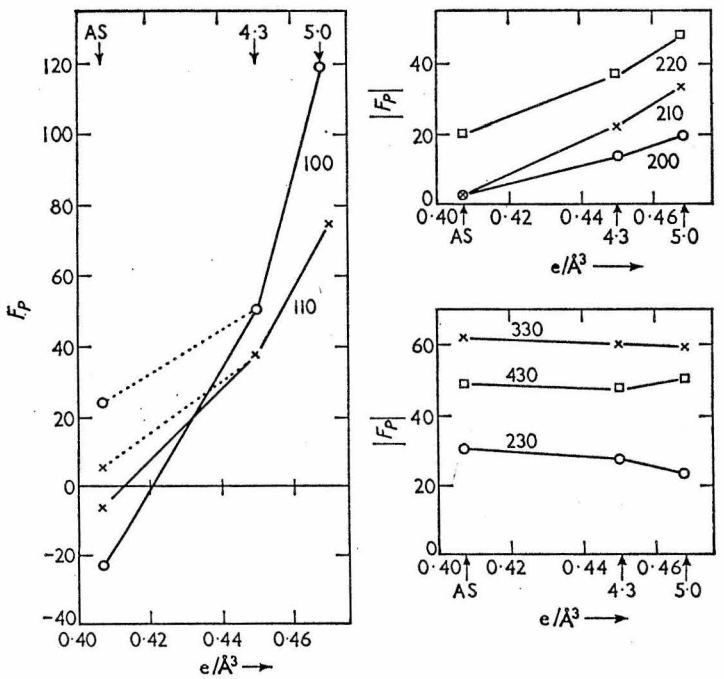



FIG. 7. Salt effect plots of $F_P$ (left) or $|F_P|$ (right) *versus* electron density of crystallizing medium.

Point AS marks the density of 95% saturated ammonium sulfate (0·407 e/Å³), 4·3 marks 4·3 M-phosphate buffer (0·45 e/Å³) and 5·0 marks 5·0 M-phosphate buffer (0·468 e/Å³). Numbers beside each curve are the (*hkl*) indices of the reflection.

then the two integrals become independent of salt density and may be treated as constants $A$ and $B$, leading to an expression linear in salt density:

$$F_{(S)} = A + B\rho_{salt}. \quad (5)$$

Plots of $|F_P|$ against salt density in Fig. 7 bear this out, and show that the change in amplitude with salt density becomes negligible around the third or fourth order from the origin. The innermost reflections are very weak in ammonium sulfate, as the contrast between molecule and medium is small. In phosphate buffer, the molecules appear in negative contrast in their very dense environment, and the inner reflections are needed to outline the molecule. The plots of $|F_P|$ against salt density for 100 and 110 are not linear, but become more nearly so if it is assumed that the signs of the reflections change in going from ammonium sulfate to phosphate. These two are the only such "cross-over" reflections found.

An independent $hk0$ sign analysis has been carried out in 4·3 M-phosphate buffer as a check, using the Pt and Hg derivatives, prepared at 2 : 1 and 10 : 1 cytochrome to metal ratios, respectively (Eisenberg & Bordner, personal communication). The structure factor data for derivatives are given in Table 1. The difference Patterson maps were essentially identical to Fig. 2(a) and (b). The only difference in behavior

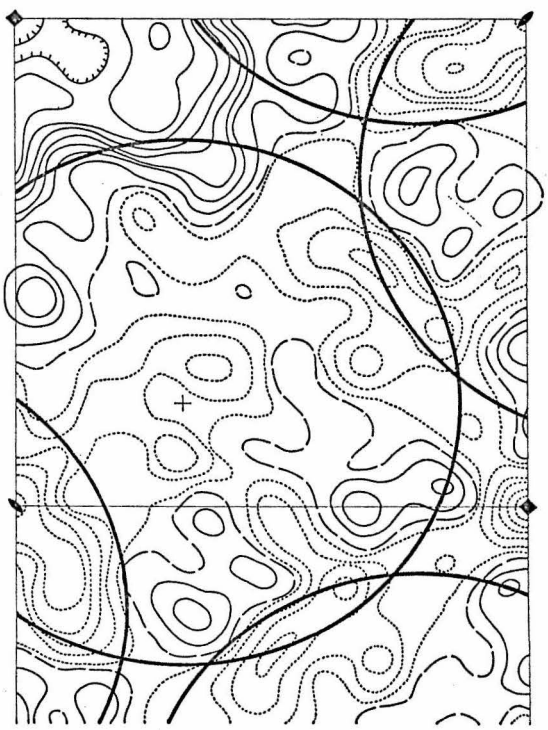


FIG. 8. Salt difference Fourier map, using as coefficients the differences between the structure factors of native cytochrome in 4·3 M-phosphate and in ammonium sulfate, each with its own proper sign.

Positive regions (solid contours) are regions of more than average accessibility to crystallizing medium in projection, and negative regions (dotted) are the least accessible to salt solution and hence are occupied by the molecule. Same axis conventions as Fig. 5.

appears to have been a 50% greater binding of Hg in phosphate over that reported for sulfate in Table 2. The only significant difference in signs determined in the two media was the reversal of the signs of the 100 and 110 reflections, in agreement with the extrapolations of Fig. 7.

If proper account is taken of the cross-over terms, then the salt difference map of Fig. 8 is obtained. This map agrees quite well with the model of the molecule as a rough sphere of diameter around 30 Å centered as originally suggested by Fig. 6. At the center of the molecule, indicated by + in Fig. 8, the salt-inaccessible region is thicker in projection than at any other place in any one molecule, and the map is most negative. As the edge of the molecule is approached going towards the origin, the thickness of molecule becomes less and the thickness of salt-accessible region increases. Finally, the region around the origin which is completely accessible to salt along the entire projected length of the *z*-axis is flat and positive. The regions where two molecules overlap in projection, around the dyad axis and between two molecules related by the screw tetrad, are forbidden to salt for a considerable length along the *z*-axis and therefore are particularly negative.

## 4. Interpretation of Results

Two lines of evidence, the Fourier map and the salt difference map, have suggested the same molecular location and diameter. Two further arguments reinforce this: molecular volume and the necessity for packing integrity of the crystal. The unit

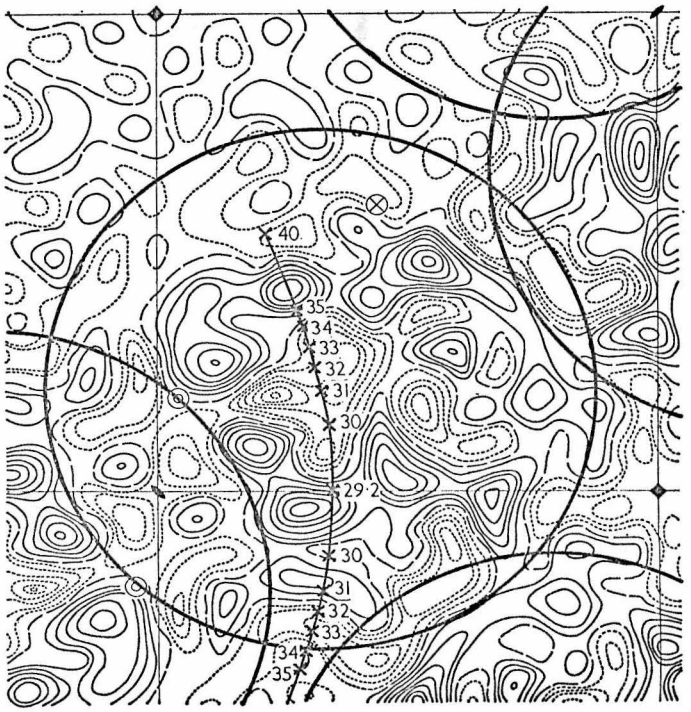



FIG. 9. Fourier projection of Fig. 6. with a sphere packing diagram superimposed, and a 31 Å diameter circle drawn about the center which is required for spheres of such diameter to be in touching contact about both the 4-fold and 2-fold screw axes.

cell in ammonium sulfate is known to be only 45% protein by weight, and in view of the similarity of mean density of protein and medium, it can be considered to be 45% protein by volume as well. If 45% of the measured cell volume is divided among four equal spheres, then each sphere must have a diameter of 31·2 Å. Moreover, if these spheres are to touch one another about both the screw dyad and the 4-fold screw tetrad, as they must to prevent collapse of the crystal, then the center of the molecule is dictated by its diameter. In Fig. 9 a sphere packing curve is superimposed on the Fourier projection. The number by each × indicates the diameter of sphere required for stable touching spheres centered at the × and at symmetry-related points in the cell. Again, it is the 31 Å sphere which best fits the region of one molecule, as originally chosen from Fig. 6. The region of overlap in projection of molecules related by the screw tetrad is negative, as a projection through a great thickness of external side chain should be. The boundaries of the molecule where overlap with neighbors does not occur are less negative. There is a "bite" out of the lower left corner of the central molecule which follows quite well the region of overlap with this molecule by the negative-density side-chains from the next molecule up the screw dyad. The unusually low density at $(\frac{1}{2}, \frac{1}{2})$ is the sole feature not explained by this model, and must represent the occupation of this tetrad axis region in three dimensions by side chains and a departure from the simple sphere model.

A sphere packing model based on the arguments of the preceding sections is shown in Plate III (a) and (b). The molecules are seen to touch at only four points, fore and aft around the screw tetrad and above and below along the screw dyad. There is considerable open space between molecules, which agrees with the observed mechanical and thermal instability of the crystals. Any compressive forces on the crystal are translated into shear forces across intermolecular contacts, and there is little mechanical resistance to deformation.

On the basis of the occurrence in most known crystal forms of cytochrome *c* of 4-fold or pseudo 4-fold X-ray pattern symmetry in a plane normal to a 35 to 45 Å crystal axis, Blow, Bodo, Rossman & Taylor (1964) proposed a common packing principle for cytochrome *c* crystal forms. The molecules are first assumed to form into screw tetrads, with a repeat of 35 to 45 Å, and these tetrad groups then are assumed to pack in various ways into the respective cells, which may or may not make full use of the $4_1$ symmetry of the subunits. This idea was shown to be consistent with several new space groups from various species (Dickerson *et al.*, in Margoliash & Schejter, 1966). It now appears that such tight screw tetrad units do exist in this structure, centered at $(\frac{1}{2}, \frac{1}{2})$, and may well exist in other crystal forms also.

Plate III (b) shows why the 240 reflection is so much more intense than any other *hk*0 reflection (Plate II(a) and Table 1). The molecules lie on the 120 planes of the crystal, yet the 120 reflection is virtually absent in ammonium sulfate. This is so because the molecules are not hard, filled spheres, but instead are hollow (in the sense of having centers of less than average density). There are two maxima instead of one when a molecule is traversed, as Fig. 9 shows, and the second-order 120 reflection, or the 240 reflection, is particularly prominent. Had the molecules not been hollow, the 120 reflection would have been stronger than the 240.

## 5. Conclusions

The conclusion from this work is that the cytochrome *c* molecule is roughly spherical in shape and approximately 31 Å in diameter. The center of the molecule is occupied

by loosely packed hydrophobic side chains. Surrounding this is a more dense framework of polypeptide chain, and outside this again is a less dense packing of hydrophilic side chains. So far, the cytochrome *c* molecule looks like a good example of the often-discussed "hydrophobic drop" structure. The packing is sensible and in agreement with the physical properties of the crystals. Confirmation of this model will have to wait upon the three-dimensional analysis now in progress.

We thank Drs David Eisenberg and Jon Bordner for permission to quote their unpublished results in phosphate buffer. Our work was supported by United States Public Health Service research grant GM-12121, the help of which is gratefully acknowledged. One of us (J. W.) is also the holder of a National Institute of Health Predoctoral Traineeship (GM-1262). This work is contribution no. 3519 of the Gates and Crellin Laboratories of the California Institute of Technology.

## REFERENCES

Ando, K., Matsubara, H. & Okunuki, K. (1965). *Proc. Japan. Acad.* **41**, 79.
Ando, K., Matsubara, H. & Okunuki, K. (1966*a*). *Biochim. biophys. Acta*, **118**, 240.
Ando, K., Matsubara, H. & Okunuki, K. (1966*b*). *Biochim. biophys. Acta*, **118**, 256.
Blow, D. M., Bodo, G., Rossmann, M. G. & Taylor, C. P. S. (1964). *J. Mol. Biol.* **8**, 606.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794.
Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G. & North, A. C. T. (1962). *Proc. Roy. Soc.* A, **265**, 161.
Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961*a*). In *Computing Methods and the Phase Problem in X-ray Crystal Analysis*, ed. by R. Pepinsky, & J. M. Robertson, p. 248. London: Pergamon Press.
Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961*b*). *Acta Cryst.* **14**, 1188.
Dickerson, R. E., Kopka, M. L., Varnum, J. C. & Weinzierl, J. E. (1967*a*). *Acta Cryst.* in the press.
Dickerson, R. E., Weinzierl, J. E. & Palmer, R. A. (1967*b*). *Acta Cryst.* in the press.
Kraut, J., Sieker, L. C., High, D. F. & Freer, S. T. (1962). *Proc. Nat. Acad. Sci., Wash.* **48**, 1417.
Lipscomb, W. N., Coppola, J. C., Hartsuck, J. A., Ludwig, M. L., Muirhead, H., Searl, J. & Steitz, T. A. (1966). *J. Mol Biol.* **19**, 423.
Margoliash, E. & Schejter, A. (1966). *Advanc. Protein Chem.* **21**, 114.
Margoliash, E. & Walasek, O. F. (1967). In *Methods in Enzymology*, ed. by S. P. Colowick & N. O. Kaplan, vol. 10. New York: Academic Press.
Sigler, P. B. & Blow, D. M. (1965). *J. Mol. Biol.* **14**, 640

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 242, No. 12, Issue of June 25, pp. 3015-3017, 1967
Printed in U.S.A.

# Location of the Heme in Horse Heart Ferricytochrome *c* by X-Ray Diffraction*

(Received for publication, April 3, 1967)

RICHARD E. DICKERSON, MARY L. KOPKA, JON WEINZIERL,‡ JOAN VARNUM, DAVID EISENBERG, AND E. MARGOLIASH

*From the Gates and Crellin Laboratories, California Institute of Technology, Pasadena, California 91109, and the Department of Molecular Biology, Abbott Laboratories, North Chicago, Illinois 60064*

## SUMMARY

**A low resolution (4 A) electron density map of horse heart cytochrome *c* crystals has been obtained by x-ray diffraction methods. The molecule is a prolate spheroid, approximately 25 × 25 × 37 A. It has a cleft or crevice along one side into which the heme is inserted, normal to the surface, with only one of the edges of the porphyrin ring exposed to solvent. The ligands occupying coordination positions 5 and 6 of the heme iron extend out from either side of the cleft. The thioether links binding the vinyl side chains of the heme to cysteinyl residues in positions 14 and 17 of the amino acid sequence are visible. One of the iron ligands can be identified from its shape and location as the imidazole side chain of the histidyl residue in position 18. The other is not an imidazole side chain and is probably located in the carboxyl-terminal half of the amino acid sequence. There is little or no α-helix present, the body of the molecule being an extended chain shell around a core of packed hydrophobic side chains.**

A considerable amount of knowledge is available with regard to the amino acid sequences, the activity, the physicochemical behavior, and the immunochemical properties of cytochromes *c* from numerous species (see Reference 1). An x-ray diffraction analysis of horse heart ferricytochrome *c* is in progress, and the present communication describes the position and environment of the heme group in the molecule as they have been revealed by the early stages of the analysis.

Horse heart cytochrome *c*, prepared by an aluminum sulfate extraction procedure (2), was crystallized in the ferric form in near saturated $(NH_4)_2SO_4$ containing 0.5 to 1.0 M NaCl. Two heavy atom isomorphous derivatives were prepared by diffusing $K_2PtCl_4$ and mersalyl (HO—Hg—$CH_2$—CH($OCH_3$)—$CH_2$—NH—CO—(*o*-$C_6H_4$)—O—$CH_2COONa$) into pregrown crystals suspended in mother liquor at pH 5.5 to 6.5. Both of these compounds bind to cytochrome *c* molecules at a single site each, and the double derivative, containing both platinum and mercury, has also been obtained. The derivatives have been characterized, the heavy atoms located, and a structure analysis in two-dimensional projection at a resolution of 4 A carried out (3).

Diffraction data for crystals of the parent protein and of the platinum and mercury derivatives have been collected in three dimensions at 4 A resolution (1475 reflections per derivative), a multiple isomorphous phase analysis has been carried out, and electron density maps have been calculated for 48 sections through the molecule, 0.86 A apart. The ratio of the mean change in $F$ produced by the heavy atom to the mean $F$ itself in the centrosymmetrical hk0 zone is 0.292 for platinum and 0.145 for mercury. The corresponding values over-all reflections are 0.201 and 0.114, respectively. The mean figure of merit with two derivatives is 0.46, and the Kraut R factors (4) are 7.8% for platinum and 6.1% for mercury. Judging from the two-dimensional analysis (Fig. 6 of Reference 3), the low figure of merit at this point is to be attributed to the use of the bare minimum of data necessary for phase determination. However, even at this stage, the individual molecules of the protein can be readily delineated and the heme group and its environment are clearly visible.

The crystals are tetragonal, space group $P4_1$, with 1 molecule of molecular weight 12,400 per asymmetrical unit (4). Cell dimensions are: $a = b = 58.45$ A, $c$ (the 4-fold axis) $= 42.35$ A. Only 45% of the crystal by volume is protein, the rest being liquid of crystallization (5). The molecules pack in a very open manner; they are grouped most tightly around a 4-fold screw axis located at ($\frac{1}{2}$, $\frac{1}{2}$, $z$), these screw assemblies then being packed so as to touch across 2-fold screw axes at ($\frac{1}{2}$, 0, $z$) and (0, $\frac{1}{2}$, $z$). There are relatively open channels through the crystal along (0, 0, $z$). Each molecule touches neighbors at only four points, fore and aft along the 4-fold screw axes, and above and below along the 2-fold screw axes (3).

Without taking into account exterior side chains, the molecule is a prolate spheroid of dimensions roughly 25 × 25 × 37 A. The mean electron density throughout the entire molecule is almost exactly equal to the electron density of the crystallizing medium, so that regions of more loosely packed van der Waals' interacting side chains appear negative on the electron density map. More dense regions with a high proportion of covalent bonds appear positive at this resolution. A photograph from the side with the heme crevice of a model built from contours at an electron density that would mark out the covalently bonded structures is given in Fig. 1. Fig. 3 is a similar photograph showing the metal-binding sites and an apparent "pseudo-channel" from the surface to the heme.

The center of the molecule is strongly negative, indicating most probably a region of packed hydrophobic side chains. Around this core is a dense, structured layer or shell, made up of extended polypeptide chain winding around to form an almost

* Contribution 3510 from the Gates and Crellin Laboratories of Chemistry. Supported by Public Health Service Research Grant GM 12121 from the National Institute of General Medical Sciences.

‡ Recipient of a National Institutes of Health predoctoral traineeship.

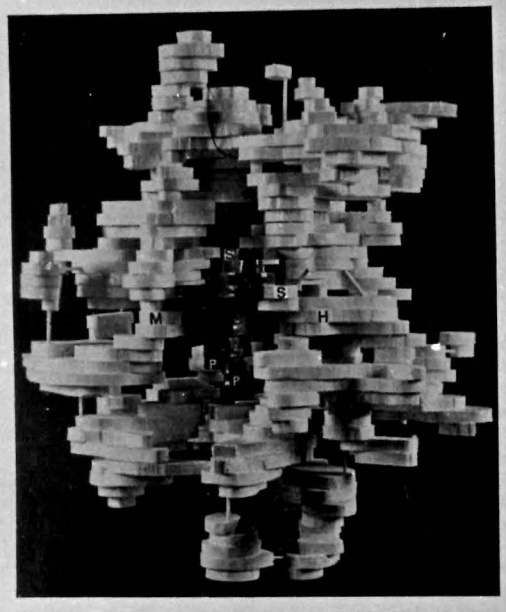



FIG. 1. A model at 4 A resolution of the molecule of horse heart ferricytochrome *c* showing the regions of the covalently bonded skeleton of the molecule. The heme crevice is visible as a vertical cleft in the *center* of this view, with the heme group painted a darker color and seen edge on. *Letters* are explained in the text.

seamless wall with three exceptions, the heme crevice and two apparent channels to the interior. Around this shell lies a more negative region, not as negative as the innermost core and probably made up of hydrophilic side chains extending into the salt region. The salt medium around and between molecules is flat and devoid of features. In comparison with hemoglobin, myoglobin, lysozyme, and ribonuclease, cytochrome *c* is the smallest protein yet studied by x-ray diffraction and is the best example of the often proposed "hydrophobic drop" model. Perhaps its simplicity is imposed upon the molecule by its size.

There are three features in the surface. The first is a crevice (6–8) along the long axis of the molecule, about 21 A long, into which the heme group is fitted (Fig. 1). The heme is parallel to the crevice and perpendicular to the surface of the molecule, with only one of its edges exposed to the medium, in general agreement with recent solvent perturbation studies of Stellwagen (9). The normal to the heme plane makes an angle of 70° with the *z* axis, in excellent agreement with the polarized absorption spectrum results of 72 ± 3° of Eaton and Hochstrasser (10) and of 65–70° of Kabat (11). The porphyrin ring extends into the interior of the molecule, which serves as a hydrophobic pocket for the heme.

The fifth and sixth coordination ligands of the iron (*H* and *M*, Fig. 1) extend from either side of the crevice, and the thioether bonds to cysteinyl residues 14 and 17 of the amino acid sequence are visible near the top of the crevice (*S* and *S*, Fig. 1). The heme propionic acid side chains (*P* and *P*, Fig. 1) do not point out into the solvent, as they do for myoglobin and hemoglobin and as has sometimes been suggested for cytochrome *c* (9). Instead, they point down toward one side of the pocket, one nearer the surface than the other, perhaps associated with one or more basic side chains within the pocket. Three sections through the electron density of the heme in a direction parallel to the *z* axis are shown superimposed in Fig. 2*A*. The relationship of the prosthetic group to the molecular surface is shown in Fig. 2*B*.

Judging from its flattened disc shape when viewed down the *z* axis, its oval (*y*, *z*) cross section and its proximity to cysteinyl residues 14 and 17, the heme iron ligand marked *H* in Fig. 1 is the imidazole side chain of histidyl residue 18 (12). On the far side of the heme, an extended polypeptide chain runs parallel to the long axis of the molecule and bears the sixth iron ligand (*M*, Fig. 1). The shape of this ligand indicates it is almost certainly not a histidyl imidazole, or a tyrosyl or tryptophyl side chain. Moreover, it is not possible for the polypeptide chain, as it can be followed on the electron density maps, to extend from one side of the heme to the other in a distance short enough for the sixth ligand to arise from histidyl residues in position 26 or 33, the only other such residues in the protein (12). Evidence that only 1 histidyl residue is concerned with hemochrome formation in cytochrome *c* has come from primary structure studies (13, 14). A tentative fitting of the amino acid sequence of horse heart cytochrome *c* (12) to the electron density contours, although necessarily of doubtful validity at the present resolution, nevertheless indicates that the sixth iron ligand most probably arises from the carboxyl-terminal half of the polypeptide chain and could readily accommodate the methionyl residue in position 80, as indicated by several lines of indirect evidence (15–21). The same tentative assignment of the polypeptide chain to the model is in remarkably good agreement with a large number of other structural implications of amino acid sequence, evolutionary, enzymic, physicochemical, and immunological data for cytochrome *c*.

The other features of the surface of the molecule are two "pseudo-channels" to the interior. These spaces are not empty channels in the true sense, and are not accessible to the solvent. These regions have a negative map density and are therefore

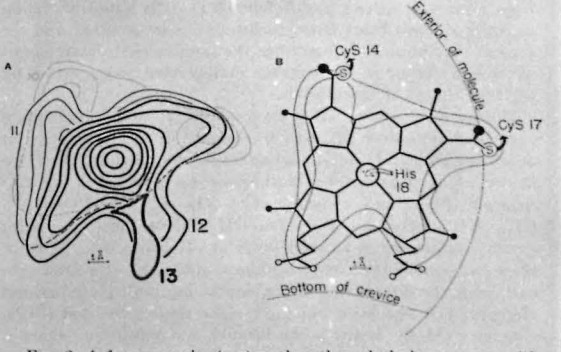



FIG. 2. *A*, four successive (*y*, *z*) sections through the heme group, with *z* vertical and *x* at 13, 12, 11 and 10/48 of a cell edge. The plane of the heme group tilts back to the *upper left*. Sections 13 and 12 contain the propionic acid side groups at the *bottom*, and section 10 shows the takeoff directions of the two thioether links to the polypeptide chain. *B*, a summary of sections 10 through 13, with an idealized heme skeleton superimposed at the same scale. The relationship of the heme to the polypeptide chain, to the bottom of its crevice and to the surface of the molecule is shown.

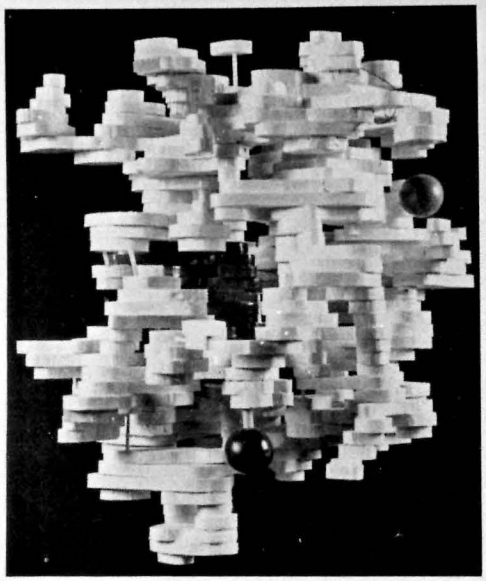


FIG. 3. A view from the right side of the molecular model of Fig. 1 showing the heavy metal-binding sites and the "pseudo-channel" to the back of the heme. The mercury site is marked by the sphere at *bottom center*, and the platinum site by the one at the *upper right*. Note the running lengths of density, believed to be extended polypeptide chain, which make up the molecular surface, the nodes of intersection where the tracing of one continuous chain becomes impossible, and the frequent complementarity of adjacent chain shapes. If side groups were added around these skeletal pathways, the molecular surface would show few gaps.

regions where loosely packed side chains in the interior meet similar side chains on the exterior, without encountering the covalently bonded shell of the molecule. One such channel extends straight down from the top of the model parallel to the $z$ axis, and the other extends inward from one side to the surface of the heme which bears the imidazole ligand, as shown in Fig. 3.

The actual length of the 104-residue polypeptide chain (104 × 3.5 A) is very near to the total running chain length observed in the model. Where two segments of the chain come close together, a node is seen at the present resolution. The precise chain paths in and out of such a node cannot be determined, making highly tentative any fitting of the amino acid sequence to the model. However, a considerable complementarity between neighboring chains can readily be detected, and if the chains were imagined to be covered with a layer of side groups, a smooth and virtually seamless wall would result. There is little or no $\alpha$-helix in the molecule. Only one location, on the side of the heme bearing the iron ligand which is not an imidazole group, could contain 1 to 1½ turns of $\alpha$-helix. Recently proposed models of cytochrome $c$ based upon the assumption of a considerable $\alpha$-helix content (22) are therefore incorrect. It is clear that the general aspects of the tertiary structure of cytochrome $c$ are quite different from those of the other heme proteins for which such information is available, myoglobin and hemoglobin.

The absolute handedness of the molecule has not yet been determined, and it is possible that Figs. 1 and 3 may represent the mirror image of the true structure. The present map represents a preliminary calculation based on the minimum amount of data needed for a multiple isomorphous replacement phase analysis. As more derivatives were added and the error level is lowered, we would expect the map to be improved but not fundamentally altered. When the third and possibly fourth derivatives have been added, the analysis will be reported in detail.

*Acknowledgments*—The authors are indebted to R. L. Gardisky and R. J. Starshak for their excellent help in the preparation of horse heart cytochrome $c$ and the search for isomorphous derivatives, and to Roberta Pratl and Kathryn Christiansen for their valuable help in collecting and processing photographic data.

REFERENCES

1. MARGOLIASH, E., AND SCHEJTER, A., *Advance. Protein Chem.*, **21**, 113 (1966).
2. MARGOLIASH, E., AND WALASEK, O. F., in S. P. COLOWICK AND N. O. KAPLAN (Editors), *Methods in enzymology, Vol. 10*, Academic Press, New York, in press.
3. DICKERSON, R. E., KOPKA, M. L., BORDERS, C. L., VARNUM, J., WEINZIERL, J., AND MARGOLIASH, E., *J. Mol. Biol.*, in press.
4. KRAUT, J., SIEKER, L. C., HIGH, D. F., AND FREER, S. T., *Proc. Nat. Acad. Sci. U. S. A.*, **48**, 1420 (1962).
5. DICKERSON, R. E., KOPKA, M. L., WEINZIERL, J., VARNUM, J., AND BORDERS, C. L., in B. CHANCE, R. W. ESTABROOK, AND T. YONETANI (Editors), *Hemes and hemoproteins*, Academic Press, New York, 1966, p. 365.
6. EHRENBERG, A., AND THEORELL, H., *Acta Chem. Scand.*, **9**, 1193 (1955).
7. GEORGE, P., AND LYSTER, R. L. J., *Proc. Nat. Acad. Sci. U. S. A.*, **44**, 1013 (1958).
8. GEORGE, P., GLAUSER, S. C., AND SCHEJTER, A., *J. Biol. Chem.*, **242**, 1690 (1967).
9. STELLWAGEN, E., *J. Biol. Chem.*, **242**, 602 (1967).
10. EATON, W. A., AND HOCHSTRASSER, R. M., *J. Chem. Phys.*, in press.
11. KABAT, D., *J. Mol. Biol.*, in press.
12. MARGOLIASH, E., SMITH, E. L., KREIL, G., AND TUPPY, H., *Nature*, **192**, 1125 (1961).
13. HELLER, J., AND SMITH, E. L., *Proc. Nat. Acad. Sci. U. S. A.*, **54**, 1621 (1965).
14. STEWART, J. W., MARGOLIASH, E., AND SHERMAN, F., *Fed. Proc.*, **25**, 647 (1966).
15. HARBURY, H. A., in B. CHANCE, R. W. ESTABROOK, AND T. YONETANI (Editors), *Hemes and hemoproteins*, Academic Press, New York, 1966, p. 391.
16. HARBURY, H. A., CRONIN, J. A., FANGER, M. W., HETTINGER, T. P., MURPHY, A. J., MYER, Y. P., AND VINOGRADOV, S. N., *Proc. Nat. Acad. Sci. U. S. A.*, **54**, 1658 (1965).
17. TSAI, H. J., AND WILLIAMS, G. R., *Can. J. Biochem.*, **43**, 1409 (1965).
18. TSAI, H. J., AND WILLIAMS, G. R., *Can. J. Biochem.*, **43**, 1995 (1965).
19. ANDO, K., MATSUBARA, H., AND OKUNUKI, K., *Proc. Japan Acad.*, **41**, 79 (1965).
20. MARGOLIASH, E., in B. CHANCE. R. W., ESTABROOK, AND T. YONETANI (Editors), *Hemes and hemoproteins*, Academic Press, New York, 1966, p. 371.
21. OKUNUKI, K., WADA, K., MATSUBARA, H., AND TAKEMORI, S., in T. E. KING, H. S. MASON, AND M. MORRISON (Editors), *Oxidases and related redox systems*, Wiley and Sons, Inc., New York, 1965, p. 549.
22. LEVINTHAL, C., *Sci. Amer.*, **214**, 42 (1966).

# An Interpretation of a Two-Derivative, 4 Å Resolution Electron Density Map of Horse Heart Ferricytochrome *c*[1]

(Received for publication, )

By R. E. Dickerson, M. L. Kopka, J. E. Weinzierl, J. C. Varnum, D. Eisenberg, and E. Margoliash*

From the Gates and Crellin Laboratories, California Institute of Technology, Pasadena, California, and the Department of Molecular Biology,* Abbott Laboratories, North Chicago, Illinois.

Running Title:

4 Å Map of Cytochrome c

# INTRODUCTION

Just eighty years ago, C. A. MacMunn published a paper (1) on the isolation and extraction from pigeon breast muscle of two new cell pigments, "histohaematin" and "myohaematin", whose reduced spectra he had observed in 1884. The disastrous dispute with Hoppe-Seyler in 1890 (2) halted work on these pigments until David Keilin began again in 1925 and coined the name "cytochrome". The identification, separation and gradual purification of cytochrome c continued in the 1930's and 40's, by Theorell and Akeson (3), Keilin and Hartree (4) and others. But it was not until 1955 that Bodo (5) finally obtained a crystalline preparation, from king penguin muscle. It took seventy years, then, for cytochrome c to progress from spectral curiosity, to crude extract, to pure protein and finally to crystalline protein, and to reach the point where people such as I would dare study it.

But crystallinity alone is not enough. The average protein crystallographer is no biochemist. For him, the protein must be not only crystalline, but easily and macroscopically crystallizable. The silken sheen which signals crystallinity to the enzymologist will drive the crystallographer to distraction. The difficulty of obtaining crystals of a size and quality suitable for x-ray work was a large factor in the abandonment of the first two attempts, by Bodo (6) and by Blow, Bodo, Rossmann and Taylor (7), to carry out a crystal structure analysis.

Older extraction methods had used conditions too drastic for the protein: trichoroacetic (4), acetic (8) or sulfuric acid (9), or organic solvents

(10). These methods were shown to partially deamidate the protein and to introduce other artifactual species which interfered with crystallization (11). The development of the gentler aluminum sulfate extraction method by Margoliash and Walasek (12) led to a preparation of a purity which made the present x-ray structure analysis possible.

## Structure Analysis Summary

The purified horse heart preparation was crystallized from 90-95% saturated ammonium sulfate solution, 0.5-1.0 M in NaCl. The molecules were found to crystallize in the tetragonal space group $P4_1$, a considerably simpler packing mode than that found previously for native cytochrome c (6, 7). In the crystal, the molecules pack around the screw axis in a spiral with a repeat every fourth molecule, 42.34 Å farther up the axis (Figure 1).[2] The screw axis units are packed in parallel with "threads" and "grooves" meshing, 58.45 Å apart. Only 45% of the crystal by weight is protein; the remainder is intermolecular liquid of crystallization.

A screening program of a fairly conventional kind was set up in an effort to find heavy atom derivatives suitable for x-ray analysis. The derivative search, data collection and subsequent structure analysis have been the subject of two recent papers (13, 14), and will only be summarized here. Two suitable heavy atom derivatives were found by diffusion trials, $PtCl_4^=$ (Pt) and mersalyl (Hg). Each of these was found to bind to the protein at a different single site without otherwise altering the folding of the protein or its manner of packing into the crystal to a detectable extent. Systematic attempts have been made to

incorporate heavy-atom-labeled carboxymethylation reagents into the protein at His33 or Met65 after the work of Okunuki et al. (see 13), but these have not yet been successful. The characterization and refinement of the heavy atom parameters and the subsequent phase analysis took place in a straightforward manner.

An electron density map has now been calculated at a resolution of 4 Å, using data from the native protein and from the Pt and the Hg derivatives. This is the theoretical minimum needed to solve the structure, and a third derivative, with the simultaneous presence of Pt and Hg on the same protein molecules, is now being added to the analysis to decrease the effect of errors. Data collection for the high resolution (2 Å) structure analysis will begin in October 1967, using these three derivatives and any more that the screening program or the carboxymethylation experiments produce. The yield of biochemically interesting information from the present two-derivative 4 Å map has been unexpectedly high, however, and it is this map which I should like to present in this Symposium.

## Overall Structure

The end product of the analysis is an electron density map, with density calculated in sections through the cell. These sections can be contoured, plotted on plexiglass sheets and stacked with the proper spacing to build up a three-dimensional map of the crystal. The problem remaining then is the one of most interest to the biochemist--that of interpreting the map, and of making meaningful chemical sense out of electron density.

With the very open crystal structure observed, individual molecules are easily separable, and in this respect cytochrome c is unusual among proteins. The molecules are prolate spheroids, ca. 25 × 25 × 37 Å, or slightly larger if exterior side groups are allowed for. Figure 1 shows a packing model of the crystal based on the electron density map, in which the molecules are represented as 31 Å diameter spheres.

The zero contour level on the map is the average electron density throughout the cell, which is almost exactly the density of the crystallizing medium between molecules. Tightly structured regions containing a high proportion of covalent bonds will appear positive on the map, and loosely packed regions with a preponderance of van der Waals contacts will appear negative. In predicting the effect of cutting off the resolution at 4 Å, it is helpful to imagine a cube, 4 Å on a side, being moved slowly through the true structure. At each instant, if the total density within the cube is averaged, and this average value is recorded at the center of the cube, the result will be very much like the detail seen in a 4 Å map.

How much of what is seen is to be trusted? How much allowance must be made for errors in the map? Certainly a map such as this one calculated with the bare minimum of derivatives needed, must be examined with caution. If either the Pt or Hg contribution to a given reflection is in error, then that reflection contributes only noise to the analysis. If for each derivative there is a 10% chance of error, then the phase analysis will be wrong for around 20% of the reflections. Fortunately, such phase errors usually lead to incoherent noise rather than cooperating to produce false detail. The map is degraded in quality but not systematically falsified. If now a third derivative is added, then two derivatives must be wrong at once for the reflection to be in error. This one new derivative reduces the number of wrong reflections to a few percent and produces a major improvement in map quality.

How much can be expected of this present map in its two-derivative stage? Many low resolution maps--lysozyme, ribonuclease, chymotrypsin--have not been interpretable in chemical terms until after the higher resolution stage. Myoglobin was an exception, because of the fact that it had two distinct substructures, the heme and the $\alpha$-helix, easily recognizable even in the 6 Å map. Cytochrome c also has such a structure, its heme group, and a stringent test of the map is the detail with which it portrays the heme.

The heme in cytochrome c is a characteristically asymmetric object (Figure 2), a flattened disk roughly 10 Å across. Its two propionic acid groups extend down in parallel, but its two thioether links to the polypeptide are oriented differently: one up, the other out to the side. The fifth and sixth iron coordination positions are both occupied by ligands attached to the

polypeptide chain, one of them adjacent to one of the thioether links.

An object of just this description is visible in the map, seen in its surroundings in Figure 3 and in closeup in Figure 4. It is a flat, irregular disk of the correct size, whose center is the most dense part of the map. A convoluted chain extends across the near side in Figure 3, and two extensions from it attach to the disk in the proper orientation. Immediately below the right connection, the chain bends back and extends another branch to the dense center of the disk. This disk obviously is the heme, with its Cys14, Cys17 and His18 connections as marked in Figure 4 and schematized in Figure 2. The two propionic acid groups extend down where expected. On the far side of the heme, a curved stretch of chain sweeps by and extends a sixth ligand to it (Figures 4,5). A final piece of evidence for identification of the heme comes from the orientation of the heme plane. The heme is not parallel to the sections shown; in Figure 4 the upper left corner is closer than the lower right. Both Eaton and Hochstrasser (15) and Kabat (16) have independently measured the angle between the heme normal and the z axis of the crystal (vertical in Figures 3-5), and have found values of 72 ± 3° and 65-70°, respectively. The angle as measured from the map is within two or three degrees of 70°.

In summary, in this test of the map, we can see an object the size of the heme as a flat disk, with side groups as small as $-CH_2CH_2COOH$ visible as projections. An extended polypeptide chain appears as a continuous rope of density. It is probable that in the rest of the map we should see the most bulky of the side chains, such as Phe, Tyr or Trp, but not differentiate between them. Charged side groups extending out from the surface of the molecule should be visible, especially if they are associated with bound ions from the solvent. The polypeptide chain should become particularly confusing

at points of close approach with another chain, whether via glycines or at bridges of any kind such as hydrogen bonds involving Thr or Ser or inter-chain salt links. We should see the total polypeptide chain, but should not expect to follow it with assurance. If there is $\alpha$-helix present, it should be clearly visible as helix with the proper parameters. What do we actually see in the map?

The center of the spheroidal molecule is strongly negative, indicating most probably a region of packed hydrophobic side chains. Around this core is a dense, structured shell, made up of ropes of density winding around to form an almost seamless wall with three exceptions, the heme crevice and two apparent channels to the interior. Around this shell lies a more negative region, not as negative as the innermost core and probably representing the loose association of hydrophilic side chains extending into the crystallizing medium. The many promontories extending into this region from the dense shell could be the skeleta of these side chains or, in some cases, bound counter ions. The salt medium around and between molecules is flat and relatively devoid of features. The total running length of ropes of density is close to the $104 \times 3.5$ Å expected for polypeptide chain, reinforcing the impression that what is observed is fully extended chain.

Figure 6 shows a solid model built from the high-density portions of the electron density map, and hence portraying the covalently bonded framework of the molecule. The most prominent surface feature is a deep crevice extending into the interior, with the heme fitted into it and connected by a ligand from each wall of the crevice. The crevice with the heme removed is shown in Figure 7. Analogous to the promontories extending out from the

surface of the molecule, there are seen to be others extending in from the inner surface of the shell towards the heme. Some of these are almost surely Phe, Tyr or Trp side groups.

The heme sits in this crevice with only one edge, that on the right in Figure 2, exposed to the solvent, in excellent agreement with the predictions of Stellwagen (17) from solvent perturbation experiments. The more exposed of the propionic groups appears to make a tenuous connection to some group on the bottom of the crevice. Pointing directly towards the buried propionic group from the back of the crevice is a bulky group (visible at the extreme right at T in Figure 5) which may be a hydrogen bonded Tyr.

The view of the molecule from the left, Figure 8, shows the heme peptide and one of the two "channels" to the center of the molecule. These are not true channels in the sense of being open to solvent; rather, they are deeply negative regions like the core, which interrupt the covalent shell. The most likely explanation is that they are regions of loosely packed side chain, perhaps hydrophobic as is the core. They may represent two hydrophobic patches on the otherwise hydrophilic surface of the molecule, and as such, may be involved in interaction with cytochrome oxidase, the cytochrome $b/c_1$ complex, or the mitochondrial membrane. The second such channel extends straight up and out the top of the molecule in Figures 7 and 8, and can be seen in Figure 14. The bottom of the molecule is closed off.

A back view of the molecule, showing the heavy atom sites, appears in Figure 9. Both Pt and Hg occupy single sites, and these sites are about as far removed from one another on the molecule as they could be, an ideal

situation for phase analysis. Chains often tend to double back on themselves or to run in parallel with nearby chains for short stretches, with complementary bends. This behavior is also found in lysozyme, chymotrypsin and ribonuclease (18, 19, 20) and has been shown in high resolution maps of these proteins to indicate cross-linking via hydrogen bonds in a manner analogous to that in the β-sheet structures of β-keratin and silks. In contrast, there is no region of the cytochrome c molecule which gives any clear indication of α-helix. Myoglobin and hemoglobin may turn out in retrospect to have been quite atypical proteins.

## The Structure of the Heme Peptide, Lys13-Thr19

The one portion of the polypeptide chain which can be identified with assurance is the run: Lys13-Cys14-Ala15-GlN16-Cys17-His18-Thr19, attached to the heme at three points. The electron density map of this region appears in Figure 10, the solid model in the same view in Figure 11, and in Figures 12 and 13, two views of a Pauling-Corey-Koltun space-filling model of the proposed chain folding. As seen in Figure 10, the chain enters from the top left, extends a promontory which makes contact with another chain across the top of the crevice, joins the heme at its upper left, bends right, then left with a very bulky region, right again in a connection to the heme at its top right corner, then left to connect with the heme at its center. From there it comes up toward the viewer, and forks down to the right and left as shown in Figure 3. The identification of these features with the known sequence is almost unavoidable, especially since the space-filling model can be oriented so as to reproduce the observed structure almost exactly.

The absolute configuration of the molecule is not yet known, and it is possible that every illustration in this paper should be reversed. The heme peptide was built with PCK models using both D- and L-amino acids, in the hope that the impossibility of one configuration or the other would provide a clue as to the absolute configuration. This did not materialize; the models are sufficiently unconstrained that a plausible matching of map density could be achieved with either form. A slight preference for the results of the D-amino acid trials led to the construction of all PCK models shown in this paper from D-amino acids and the marginal expectation that the high resolution work will shown the molecule in need of inversion. This inversion was made in the abstract of this Symposium paper and in the J. Biol. Chem. paper (14), but the present convention will be maintained until the high resolution analysis settles the question.

If this interpretation of the heme peptide is correct, then Lys13 plays a critical role in holding the molecule together. As Figure 14 shows, the bridge across the top of the crevice is quite tenuous. It is easy to imagine that, in the absence of a group similar to Lys13, the two halves of the molecule could fall open and the heme be exposed. Interestingly enough, locus 13 is evolutionarily very conservative. As Table I shows, this site is Lys in every vertebrate species studied, and is Arg in most invertebrates and microorganisms. Pseudomonas has Gly at this point but has a Lys immediately preceding it.

Ala14 appears to be exposed to the solvent, and alanine is one of the few hydrocarbon side groups which in other proteins has been found on the

outside of the molecule to an appreciable extent (19,21). In other vertebrates, it can be replaced by Ser, of similar bulk and even more compatible with a polar environment. Both yeast isomers have Glu at this point, acceptable in a polar medium.

GlN16 plays an enigmatic role. It extends to the left in Figures 10 and 11, and possibly forms a hydrogen bond with the main chain just before Lys13. It is conserved in all vertebrates, but is replaced by Glu in Neurospora and Candida, and Leu in yeast Iso-1. In the vertebrates, it is tempting to think that its function is to double the chain back on itself and to give structure to that part of the chain to which the heme must attach.

Cys14, Cys17 and His18 are absolutely conserved, and their function is apparent. The fork visible in the map at Thr19 may be vital to the subsequent folding of the molecule. Threonine is present in all species studied except Neurospora, which has Gly. But even Neurospora has a potential chain-forking residue at the next position, Glu20, and this could represent a minor local alteration in configuration.

## Possible Interpretations of Other Parts of the Molecule, and a Mechanism for Chain Folding at the Ribosome

The discussion to this point has been reasonably conservative, and the conclusions have a high probability of correctness. But one of the prime opportunities of a Symposium such as this is that of presenting tentative conclusions in the hope of provoking a discussion. There are several other suggestive features of the structure as it now stands which I should like to

present as a series of hypotheses. I shall try to make them plausible, and would be surprised if they were totally wrong, but cannot claim that they are on anything like the firm ground of the previous sections.

Hypothesis 1—A structure for the N-terminal end of the polypeptide chain, and a mechanism of folding of the molecule at the ribosome.

If Lys13 is located correctly, then where is the section Gly1--GlN12? At Lys13 the chain could go back over the top of the crevice in Figure 3, or to the left into the region with the "Val11" label. The tenuousness of the crevice-top connection and the solidity of the path to the left make the latter more likely. This path leads to a very tightly structured region, almost an entity to itself, visible at the upper left in Figures 3 and 8 and in closeup in Figure 15. Continuing around the left corner of Figure 15, there appears a loop in which the chain is doubled back upon itself and cross-connected (Figure 16). The lower leg of the loop returns to the side visible in Figure 15 and either ends, or else joins an adjacent horseshoe bend of chain coming up from below in a T joint. It is more likely that this rather tenuous join is only a close contact between chains, and that the chain in question does come to an end at the point marked "Ac" for "acylated end" in Figure 15.

The N-terminal dodecapeptide contains a characteristic pattern of charged groups, glycines and hydrophobic groups, completely invariant among all the vertebrates studied (Table I). Particularly noteworthy are the negative charges in one half, matched by positive charges in the other, with Gly6 in between. Several attempts were made to build this peptide from space-filling models, trying various bridging schemes and seeking to match the details of this region on the map. A very plausible model can be

prepared, using the bridging scheme of Figure 17. In this model, the hairpin bend in the chain is initiated by the interaction of the adjacent Glu4/Lys5 charged side groups and facilitated by the presence of Gly6. It is then locked into place by the interaction of the side chain of Asp2 with the other chain, either a salt link or a hydrogen bond. In some invertebrates, Asp2 is replaced by AsN or Ser, incapable of salt links but quite capable of forming hydrogen bonds. This may indicate that in vertebrates the principal if not the only factor in bridging is the formation of a hydrogen bond. The backbone of the polypeptide chain is not forced into an unusual or strained configuration, and in many regions resembles a fully-extended, alternating side group chain.

If this cross-linked hairpin is then given a right-angle bend at Lys7.. Asp2 to agree with the electron density map, the two remaining Lys side chains can extend out into the solvent, the side chains of Val3, Ilu9, Phe10, and the hydrocarbon part of the Lys5 side chain all pack into the inside of the fold, and Val11 is brought near the acylated N-terminus of the chain, next to the back-of-the-heme channel visible in Figure 8.

This interpretation of the map is marked by labels in Figure 18, in a view identical to that of the wooden model in Figure 15. Figure 19 shows a similar view of a PCK model of this folding scheme. The initial loop in the chain, cross-linked by Asp2, is shown as a PCK model in Figure 20 in the same orientation as the wooden model in Figure 16. Note in Figure 16 the two projections to the top right, suggestive of Lys side chains, and the two close approaches of neighboring chains at glycines: the horseshoe bend coming up from the bottom and approaching Gly1, and the close contact near Gly6.

Figure 21 shows the Glu4/Lys5 side chain interaction as seen from within the molecule in a direction opposite to that of Figure 18, and Figure 22 shows a PCK model of the same region. Note the clustering of non-polar groups on the inner (left) surface.

The entire N-terminal dodecapeptide as constructed forms a relatively rigid structure, bent like a cupped hand with the palm lined with hydrophobic side chains. With the evidence of Margoliash (22) that the heme is attached to the polypeptide chain as the chain comes off the ribosome rather than to a complete, folded protocytochrome, a self-consistent picture appears as to how the molecule might fold upon synthesis:

1) As the first dozen residues are synthesized at the ribosome, the Glu4/Lys5 side chain interaction bends the chain back upon itself to form a loop, aided by Gly6 and locked into place by Asp2. This loop is cupped and its stability strengthened by the hydrophobic interactions of Val3, Ilu9, and Phe10 at one place, and by the interaction of Val11 and the acylated N-terminus. A charged, unacylated N-terminus here is structurally unacceptable. (There is an interesting evolutionary sidelight here. The invertebrates, which have four to nine additional residues on the N-terminal end, tend to have a charged group, Lys, in place of valine.)

2) As residues Lys13-Thr19 come off the ribosome, the doubling back of the chain at GlN16 brings the chain into roughly the proper shape for interacting with the heme. It is possible that the hydrophobic region of the initial dodecapeptide interacts with the attaching enzyme or helps to hold the heme during attachment.

3) After the attachment of the heme, the heme and its peptide itself form a nucleus for the subsequent folding of the molecule. As a hydrophobic

region of the molecule comes off the ribosome, it tends to fold around the heme, bringing the chain back and maintaining a compact molecule. It is this tendency of the non-polar regions to surround the heme which is responsible for the observed shell structure one chain layer thick, and the creation of the molecule in the shape of a hand folded about the heme. The necessity for hydrophobic interactions for the proper folding of the molecule gives rise to the evolutionary conservatism of the non-polar residues.

4) At one point in the folding, a loop of chain comes close enough to the evolutionarily conserved Lys13 (or Arg13 in invertebrates) to close off the top of the crevice and add to the stability of the molecule.

This model may be wrong in many details. The essentials of the model: the doubling back of the N-terminal dodecapeptide as a nucleus for heme attachment and the use of the heme itself as a template for subsequent folding of the molecule, seem plausible at least to the author, and may be correct in outline although false in the precise form presented here.

Hypothesis 2--The location on the map of Trp59-Tyr67, and the possible identification of the sixth iron ligand.

Just below the back-of-the-heme channel in Figure 8 begins a long run of what appears to be a single extended polypeptide chain. It runs up near Gly1, back and down again, bends left in Figure 9, runs to the far corner of the molecule, doubles sharply back upon itself and rises up and over the shoulder of the molecule. Part of this chain is shown in more detail in Figure 23. The chain then continues up and over the "right side" of the

molecule (Figure 7) and down past the heme at the sixth ligand site. From there its path becomes uncertain.

The shapes of several of the side groups along this chain are particularly interesting. Where the first loop bends left at the bottom in Figure 9, there appears a flat, bulky side group pointing in to the interior of the molecule and the heme (P in Figure 23). It has the unsymmetrical shape of Trp, but might also be the Phe or Tyr. Immediately beyond it, a long promontory extends from the polypeptide chain into the solvent (L in Figure 23). This is probably a charged side chain, and from its length is more likely Lys or Arg than Asp or Glu. The sharp doubling back (T in Figure 23) occurs at a point of close approach to another chain, one of the typical ambiguous nodes where the connections between incoming and outgoing chains is not sure. The choice indicated has been made on the basis of the total chain fitting scheme. If this is a point of cross-linking of two chains, then the evidence of lysozyme suggests possibly Ser or Thr at this point (18). Beyond this node, at R in Figure 23, is another aromatic side chain pointing to the interior, and by shape more likely Phe or Tyr than Trp. Most significantly, it points directly at the more buried of the two heme propionic acid groups (T in Figure 5). In ribonuclease (20), three tyrosines lie near the surface with hydroxyl groups exposed. Three more wholly or partially buried ones, however, all have their hydroxyl groups oriented in such a way as to suggest hydrogen bonding--two to main chain carbonyls and one to the carboxyl of an Asp. Similar Tyr hydrogen bonding has been observed in lysozyme and myoglobin as well. It is possible that such a "neutralization" of the polar group is necessary in order for a Tyr to exist in a strongly hydrophobic environment. From another

viewpoint, the provision of an aromatic group with some polar hydrogen bonding character may be the method of accommodating a propionic acid side chain within a non-polar milieu.

Is it possible to find a stretch of sequence with the following characteristics:

(TRP, Phe, Tyr); (LYS, ARG, Asp, Glu) . . . . (Ser, Thr) . . . .
(TYR, Phe, Trp) ?

(The choices within parentheses represent alternatives for a single residue.) The most promising fit is the sequence: -Thr58-Trp59-Lys60-Glu61-Glu62-Thr63-Leu64-Met65-Glu66-Tyr67-Leu68-Glu69-. If this sequence is assumed for the chain between P and R in Figure 23, then the shapes of all four lettered side groups are accounted for. This model brings four glutamic acids into a cluster at one locus on the surface of the molecule, which, if correct, may be functionally significant. It accounts for the extreme evolutionary conservatism of the Thr63--it is needed to hold the molecule together at its back corner.[3] It explains the accommodation of the heme propionic group in a non-polar environment, and accounts for one of the two buried tyrosines. And most interesting of all, if the chain is continued up over the shoulder of the molecule as shown in Figure 23 and then down past the heme (Figure 7), it arrives at the heme just in time to bring Met80 to the sixth ligand site.

Hypothesis 3--The cytochrome oxidase binding site.

The model proposed in Hypothesis 2 brings Lys 72 and Lys 73 to the upper part of the right face of the molecule. Moreover, although the chain

path cannot be followed with assurance past residue 82 or so, the lysine triplet Lys86-88 must be somewhere nearby, perhaps on the lower half of this right face. Of these five cationic groups, all but Lys88 are evolutionarily invariant. It is proposed that it is this positively charged right face of the molecule which serves as the binding site for cytochrome oxidase.

Cytochrome oxidase is an acidic protein, and is known to interact with cytochrome c at least in part by electrostatic forces (11). Okunuki (23) has shown that the blocking of either Lys72 or Lys73 with a trinitrophenyl group inhibits the reaction of cytochrome c with its oxidase. Margoliash (11) has suggested that the evolutionarily constant segment Asn70-Met80 which bears these two lysines, may be constant just because it is necessary to form the oxidase binding site. If the right side of the molecule is this binding site, then the heme in its cleft would be quite accessible for direct electron transfer. The termini of the two pseudochannels on the surface of the molecule are about as far removed from this hypothetical binding site as they could be. If these are in fact hydrophobic "patches" on the surface, it is possible that they may be involved in binding in some manner to the mitochondrial cell wall in such a way as to present the binding site and the heme to the oxidase.

Hypothesis 4--A possible conformational change upon reduction of the heme.

Several differences in physical and chemical properties have been found between oxidized and reduced cytochrome c which suggest a somewhat tighter folding of ferrocytochrome. These are summarized in (11), and include differences in spreading behavior at air-water interfaces, greater

thermal stability and resistance to proteolytic digestion in the reduced form, nuclear magnetic resonance and optical rotatory dispersion spectra differences, and the diminished tendency of the ferrocytochrome heme to complex with other ligands. The behavior of reduced cytochrome c on elution from cation-exchanger columns is compatible with the loss or burial of one positively charged group.

The heme peptide, containing the His18 ligand, is seen from the x-ray work to be a tight, compact structure, with little possibility for mechanical variance. The sixth ligand, in contrast, is carried by a free-standing loop of chain which from Figure 7 does not appear to give any structural integrity to the right side of the cleft.

One possible configuration change upon reduction might be the folding inward of the heme and this flexible chain to form a more compact object and to close off the opening of the crevice. This might even involve a breaking of the heme ligand to this chain and a reforming of the sixth bond with another donor buried deeper inside the crevice. Such a localized conformation change could explain most of the changes mentioned above. It might bury Lys79, just above the ligand, and explain the cation-exchange column observations. It would certainly make the heme less accessible to foreign ligands. If the picture which has been drawn of the molecule as tenuously held together at the top of the crevice and liable to unfolding is correct, then the closing in of the crevice would make the molecule more stable to heat and probably to proteolytic digestion as well. And finally, it appears that the residue just above the sixth ligand residue, which has been proposed to be Lys79, makes a close contact with a neighboring molecule up the screw axis in the crystal.

If, in the reduced form, this charged group were removed from the exterior and this presumed salt link were made impossible, it would then be reasonable that crystallization would become much more difficult, as has actually been observed.

## Conclusions

At present, some structural features are known with a fair degree of assurance from the two-derivative, 4 Å map. These include the overall lay-out of the molecule, the geometry of the heme and its peptide, and the nature of the crevice. Other features are only hypothesized, however, and the two categories should not be confused. It is likely that we know something about the folding of the N-terminal end of the chain, and the chain-folding mechanism is at least plausible. The mapping of the chain across the back of the molecule, and the interpretation of the right side are about the best guesses which can be made on the basis of the present map and what is known about the chemical and physical properties of the protein. Whether all of these hypotheses are examples of the power of a 4 Å map or illustrations of enthusiastic self-deception on the part of the observers will have to be decided at higher resolution. Although they look plausible and tie a considerable amount of chemical information together, the safest judgment upon them at this point is: "Se non è vero, è bene trovato".

## Summary

The gross features of the structure of the molecule of horse heart ferricytochrome c, obtained from an electron density map at 4 Å resolution using Pt and Hg derivatives, is illustrated with stereo photos of the plexiglass map and of plywood and PCK models. The interpretation of J. Biol. Chem., 242, 3015 (1967) is amplified as to the overall architecture of the molecule and the structure of the heme peptide. Four hypotheses are presented, based upon the present model and the available chemical and physical evidence, to be tested against the high resolution structure analysis and the subsequent analysis of ferrocytochrome:

1) A structure for the N-terminal dodecapeptide, and a mechanism of folding of the molecule at the ribosome.

2) The location on the molecule of Trp59-Tyr67, and the possible identification of the sixth iron ligand.

3) A possible cytochrome oxidase binding site.

4) A possible conformational change upon reduction of the heme.

## Footnotes

[1] Contribution No. 3524 from the Gates and Crellin Laboratories of Chemistry. Supported by United States Public Health Service Research Grant GM-12121 from the National Institutes of General Medical Sciences. One of the authors (J. W.) is also the holder of a National Institutes of Health predoctoral traineeship.

[2] The photographs in this paper are stereo pairs, positioned for viewing either with the unaided eye or with a conventional binocular hand viewer such as the CF-8 Stereoscope obtainable from Abrams Instrument Corporation, Lansing, Michigan, U.S.A. The following procedure will enable many people to see the photographs in depth without a viewer.

a) Hold the page at arms' length.

b) Look past the page; relax the eyes until two images of each half of the stereo photo are seen. It is sometimes helpful to look at a distant object over the edge of the page and then shift the page into view without readjusting the convergence of the eyes.

c) Adjust the degree of convergence of the eyes until the center two of the four half-images coincide. This is the proper convergence for stereo vision. If the page was held initially at arms' length, the merged stereo image should be in focus. If your eyes and the stereo pair are not parallel, then registry of the two half-images is impossible. If registry is not initially achieve rotate the page clockwise and counterclockwise by a few degrees until the stereo image snaps into place.

d) Bring the page towards your eyes slowly until the retention of near-

object focussing and distant-object convergence becomes uncomfortable. With practice, the image can be held at normal reading distance.

The photographs were taken with a Steritar B attachment for the Zeiss Contaflex, which places the two stereo images on the two halves of one 35mm frame of the negative. The film used was Eastman Kodak Plus-X, speed ASA 125.

[3] The entire sequence: Ala51-----Ilu81 is almost invariant throughout all the vertebrates, with the exception of such non-radical substitutions as Ser for AsN54, Val or Ilu for Thr58, Gly generally for Lys60 of horse, and Asp for Glu62. This is the conservatism which would be expected for a long chain which formed one of the principal structural entities of the molecule.

## Figures

Figure 1: Packing of idealized cytochrome c molecules into the crystal. The fourfold screw axes described in the text come up to the left out of the photograph. Note the large open channel down the center of the four-screw fragment of crystal shown here.

Figure 2: The heme peptide, the arrangement of side groups on the heme, and the connections between heme and peptide as revealed by x-ray analysis.

Figure 3: Sections through the three-dimensional electron density map at 4 Å resolution, showing the "left" side of the molecule, the heme within, and the pseudochannel down to it. The round paper labels identify postulated side chain regions, and the smaller dots, rough $\alpha$-carbon positions. The placing of Lys13-Thr19 is reasonably correct; that of Gly1-GlN12 is hypothetical only. The direction of the chain fork at Thr19 is unknown. Sections nearest the observer have been removed. Contours are at intervals of fifty on an arbitrary scale, with the zero contour being the average density throughout the cell. Only positive contours are plotted on this map.

Figure 4: The heme and its immediate environment. The Cys14, Cys17 and His18 connections are marked. Two propionic acid groups extend down below the heme. The exterior of the molecule is to the right of the heme, and the center of the molecule is to its left. The density at the center of the heme iron is 480 on the arbitrary scale, while that along extended polypeptide

chains is usually 130-280. The most dense part of the map other than the heme center is the bottom of the N-terminal loop (Figure 16), with a maximum of 353.

Figure 5: The right side of the heme, and in front, the extended chain which bears the sixth iron ligand and forms the right side of the heme crevice (also visible behind the heme in Figure 4). Several sections closest to the viewer have been removed. Note the possible interaction of the leftmost propionic acid group with something at the bottom of the crevice.

Figure 6: A model of the molecule taken from the electron density map of the previous Figures. The contour selected for display is halfway between the first and second contours of the plexiglass map. Note the heme in its crevice, the heme peptide forming the left side of the crevice and the curved extended chain of Figure 5 on the right. X' marks a close contact with the next molecule up the fourfold screw axis, and X marks the equivalent contact with the next molecule down the axis. If the sixth iron ligand is Met80 (for which there is as yet no x-ray evidence one way or the other), then the close intermolecular contact marked by X' would involve a positive charge, Lys79.

Figure 7: Model with the heme removed. S and S' mark the thioether links to Cys17 and Cys14, respectively, H marks His18, and L marks the interaction of Lys13 across the top of the crevice. Note the pronounced cavity within the molecule, and the several bulky groups which appear to extend inward from the cavity walls toward the heme.

Figure 8: The left side of the molecule, showing the heme peptide and the pseudochannel to the back of the heme.

Figure 9: The back of the molecule, showing the Pt site (+), the mersalyl Hg site (□), the proposed N-terminal loop at the upper right, and a long stretch tentatively believed to be a continuous polypeptide chain path (see Figure 23).

Figure 10: The polypeptide chain on the His18 side of the heme. Probable side group loci and possible main chain $\alpha$-carbon sites are marked as in Figure 3. Note the close approach of the next molecule up the fourfold screw axis near another positively charged group, Lys13, and Ala15 (X).

Figure 11: Closeup of Figure 8, showing the heme peptide, the back-of-the-heme channel and the chain fork at Thr19 (T). Compare Figure 10.

Figure 12: PCK space-filling model of the heme group and its peptide Lys13-Thr19. Visible side groups are numbered as in Figure 2.

Figure 13: View of the heme and its peptide from the interior of the crevice.

Figure 14: View down on the top of the molecule from the right front. L marks the postulated Lys13 bridge across the top of the crevice. The top of the heme is visible at the bottom. Note a second pseudochannel down into the molecule from the top. The region at upper center, behind this channel, is postulated to be the N-terminal Gly1-GlN12 region (see text).

Figure 15: The proposed N-terminal peptide, with side group loci numbered as in Figure 17. Same view of the moleule as Figure 8.

Figure 16: The N-terminal loop with its bridge, seen from the back of the molecule, as in Figure 9.

Figure 17: The chain folding scheme which is proposed for Gly1-GlN12 to explain the observed electron density, and which is postulated as a nucleus for the folding of the polypeptide chain around the heme as the molecule is formed. [H]— represents a hydrophobic side chain, and ⊕— or ⊖—, a charged group.

Figure 18: The N-terminal peptide region of the electron density map, with side chains marked as in Figure 3. Sections containing the heme, at the extreme rear, have been removed for increased illumination. Compare Figure 3, in which the nearer sections have been removed instead.

Figure 19: PCK model of the Gly1-Thr19 peptide in the same orientation as Figures 15 and 18.

Figure 20: PCK model of the N-terminal loop of Figure 16.

Figure 21: The postulated Glu4/Lys5 salt link seen from within the molecule in a direction opposite to that of Figure 18.

Figure 22: PCK model of the salt link of Figure 21, represented for ease of construction as a hydrogen bond.

Figure 23: The lower right rear of the molecule, showing a chain which may contain residues Trp59-Tyr67 (see text).

## References

1. MacMunn, C.A., J. Physiol., 8, 51 (1887).
2. See the excellent review of this dispute in: Keilin, David, "The History of Cell Respiration and Cytochrome", Chapter 6, Cambridge (1966).
3. Theorell, H., and Akeson, A., J. Am. Chem. Soc., 63, 1804, 1812, 1818 (1941).
4. Keilin, D., and Hartree, E.F., Biochem. J., 39, 289 (1945).
5. Bodo, G., Nature, 176, 829 (1955).
6. Bodo, G., Biochim. Biophys. Acta, 25, 428 (1957).
7. Blow, D.M., Bodo, G., Rossmann, M.G., and Taylor, C.P.S., J. Mol. Biol., 8, 606 (1964).
8. Hagihara, B., Okunuki, K., et al., J. Biochem (Tokyo), 45, 551, 565, 725 (1958).
9. Paleus, S., Acta Chem. Scand., 14, 1743 (1960).
10. Morrison, M., Hollocher, T., Murray, R., Marinetti, G., and Stotz, E., Biochim. Biophys. Acta, 41, 334 (1962).
11. Margoliash, E., and Schejter, A., Adv. Prot. Chem., 21, 114 (1966).
12. Margoliash, E., and Walasek, O.F., in "Methods in Enzymology", (S.P. Colawick and N.O. Kaplan, eds.), Vol. 10, Acad. Press (1967).
13. Dickerson, R.E., Kopka, M.L., Borders, C.L., Jr., Varnum, J.C., Weinzierl, J.E., and Margoliash, E., J. Mol. Biol. (1967), in press.
14. Dickerson, R.E., Kopka, M.L., Weinzierl, J.E., Varnum, J.C., Eisenberg, D., and Margoliash, E., J. Biol. Chem., 242, 3015 (1967).
15. Eaton, W.A., and Hochstrasser, R.M., J. Chem. Phys., (1967), in press

16. Kabat, D. (1967), private communication.

17. Stellwagen, E., J. Biol. Chem., 242, 602 (1967).

18. Blake, C.C.F., Mair, G.A., North, A.C.T., Phillips, D.C., and Sarma, V.R., Proc. Roy. Soc., B167, 365 (1967).

19. Matthews, B.W., Sigler, P.B., Henderson, R., and Blow, D.M., Nature, 214, 652 (1967).

20. Wyckoff, H.W., Hardman, K.D., Allewell, N.M., Inagami, T., Johnson, L.N., and Richards, F.M., J. Biol. Chem., (1967), in press.

21. Kendrew, J.C., Brook. Sym. Biol., 15, 216 (1962).

22. Margoliash, E., unpublished work.

23. Okunuki, K., Wada, K., Matsubara, H., and Takemori, S., in "Oxidases and Related Redox Systems", (T.E. King, H.S. Mason, M. Morrison, eds.), p. 549. Witey, New York, 1965.

Table I

Cytochrome c Sequence Comparisons

from the N-Terminal End Through the Heme Attachment

| Residue: | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | -2 |
|---|---|---|---|---|---|---|---|---|---|
| Vertebrates | | | | | | | | | |
| Man[3] | | | | | | | | | |
| Horse[4] | | | | | | | | | |
| Chicken[5] | | | | | | | | | |
| Pekin Duck | | | | | | | | | |
| Rattlesnake | | | | | | | | | |
| Tuna | | | | | | | | | |
| Ancestral (?) | | | | | | | | | |
| Invertebrates | | | | | | | | | |
| S. Cynthia | | | | | | $H_2N$ — Gly | | Val | Pr |
| Neurospora | | | | | | $H_2N$ — Gly | | Phe | Se |
| Yeast Iso-1 | | | | | $H_2N$ — Thr | | Glu | Phe | Ly |
| Candida Krusei | | | | $H_2N$ — Pro | | Ala | Pro | Phe | Gl |
| Wheat germ | | Ac[2] — Ala | | Ser | Phe | Ser | Glu | Ala | Pr |
| Pseudomonas | | | | | | | | | |

35

| -1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Ac — | Gly | Asp | Val | Glu | Lys | Gly | Lys | Lys | Ilu | Phe |
| Ac — | Gly | Asp | Val | Glu | Lys | Gly | Lys | Lys | Ilu | Phe |
| Ac — | Gly | Asp | Ilu | Glu | Lys | Gly | Lys | Lys | Ilu | Phe |
| Ac — | Gly | Asp | Val | Glu | Lys | Gly | Lys | Lys | Ilu | Phe |
| Ac — | Gly | Asp | Val | Glu | Lys | Gly | Lys | Lys | Ilu | Phe |
| Ac — | Gly | Asp | Val | Ala | Lys | Gly | Lys | Lys | Thr | Phe |
| Ac — | Gly | Asp | Val | Glu | Lys | Gly | Lys | Lys | Ilu | Phe |
| | | ⊖ | H | ⊖ | ⊕ | | ⊕ | ⊕ | H | H |
| Ala | Gly | AsN | Ala | Glu | AsN | Gly | Lys | Lys | Ilu | Phe |
| Ala | Gly | Asp | Ser | Lys | Lys | Gly | Ala | AsN | Leu | Phe |
| Ala | Gly | Ser | Ala | Lys | Lys | Gly | Ala | Thr | Leu | Phe |
| GlN | Gly | Ser | Ala | Lys | Lys | Gly | Ala | Thr | Leu | Phe |
| Pro | Gly | AsN | Pro | Asp | Ala | Gly | Ala | Lys | Ilu | Phe |
| | | $H_2N$ — | Glu | Asp | Pro | Glu | Val | Leu | Phe | Lys |

| | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|
| K.Wada $C_1$: | | | Lys | Cys | ? | ? | Cys | His | Thr | Val |
| | Ilu | Met | Lys | Cys | Ser | GlN | Cys | His | Thr | Val |
| | Val | GlN | Lys | Cys | Ala | GlN | Cys | His | Thr | Val |
| | Val | GlN | Lys | Cys | Ser | GlN | Cys | His | Thr | Val |
| | Val | GlN | Lys | Cys | Ser | GlN | Cys | His | Thr | Val |
| | Ilu | Thr | Lys | Cys | Ser | GlN | Cys | His | Thr | Val |
| | Val | GlN | Lys | Cys | Ala | GlN | Cys | His | Thr | Val |
| | Val | GlN | Lys | Cys | Ala | GlN | Cys | His | Thr | Val |
| | H | | ⊕ | (heme) | | | (heme) | (heme) | | H |
| | Val | GlN | Arg | Cys | Ala | GlN | Cys | His | Thr | Val |
| | Lys | Thr | Arg | Cys | Ala | Glu | Cys | His | Gly | Glu |
| | Lys | Thr | Arg | Cys | Glu | Leu | Cys | His | Thr | Val |
| | Lys | Thr | Arg | Cys | Ala | Glu | Cys | His | Thr | Ilu |
| | Lys | Thr | Lys | Cys | Ala | GlN | Cys | His | Thr | Val |
| | AsN | Lys | Gly | Cys | Val | Ala | Cys | His | Ala | Ilu |

[1] Indicates free terminal amino group.

[2] Indicates acylated terminal amino: $CH_3$-CO-NH-$CHR_1$-....

[3] Includes man, rhesus monkey, chimpanzee.

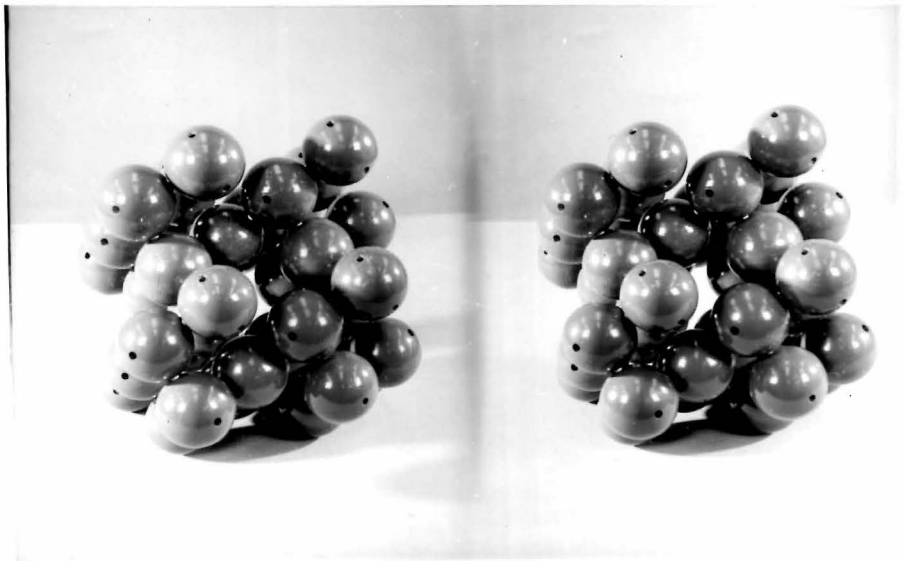
[4] Includes horse, cow, pig, sheep, dog, rabbit, kangaroo, snapping turtle.

[5] Includes chicken, turkey, king penguin, pigeon.

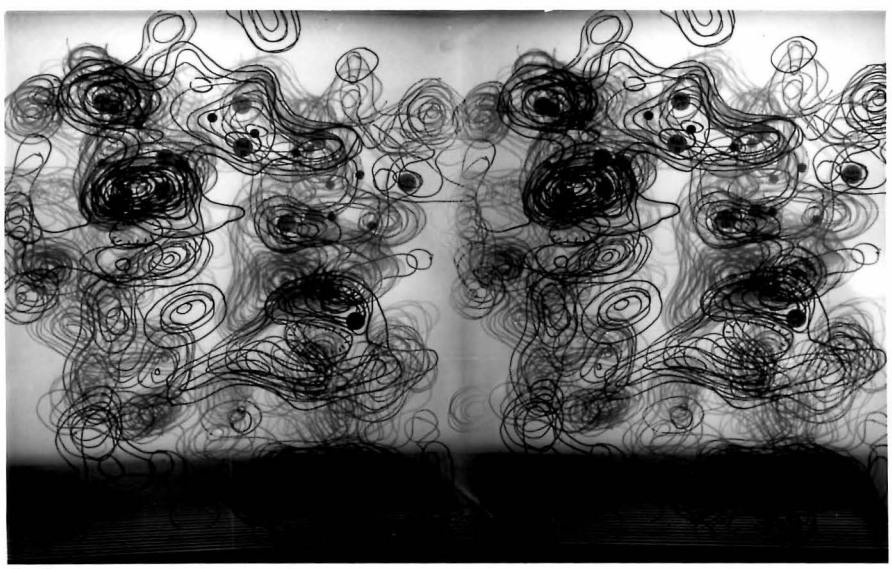
Sequence information has been taken from the Atlas of Protein Sequence, 1966 (R. V. Eck and M. O. Dayhoff, eds.), with the addition of the following:

1) Chimpanzee, pigeon : E. Margoliash, private communication.

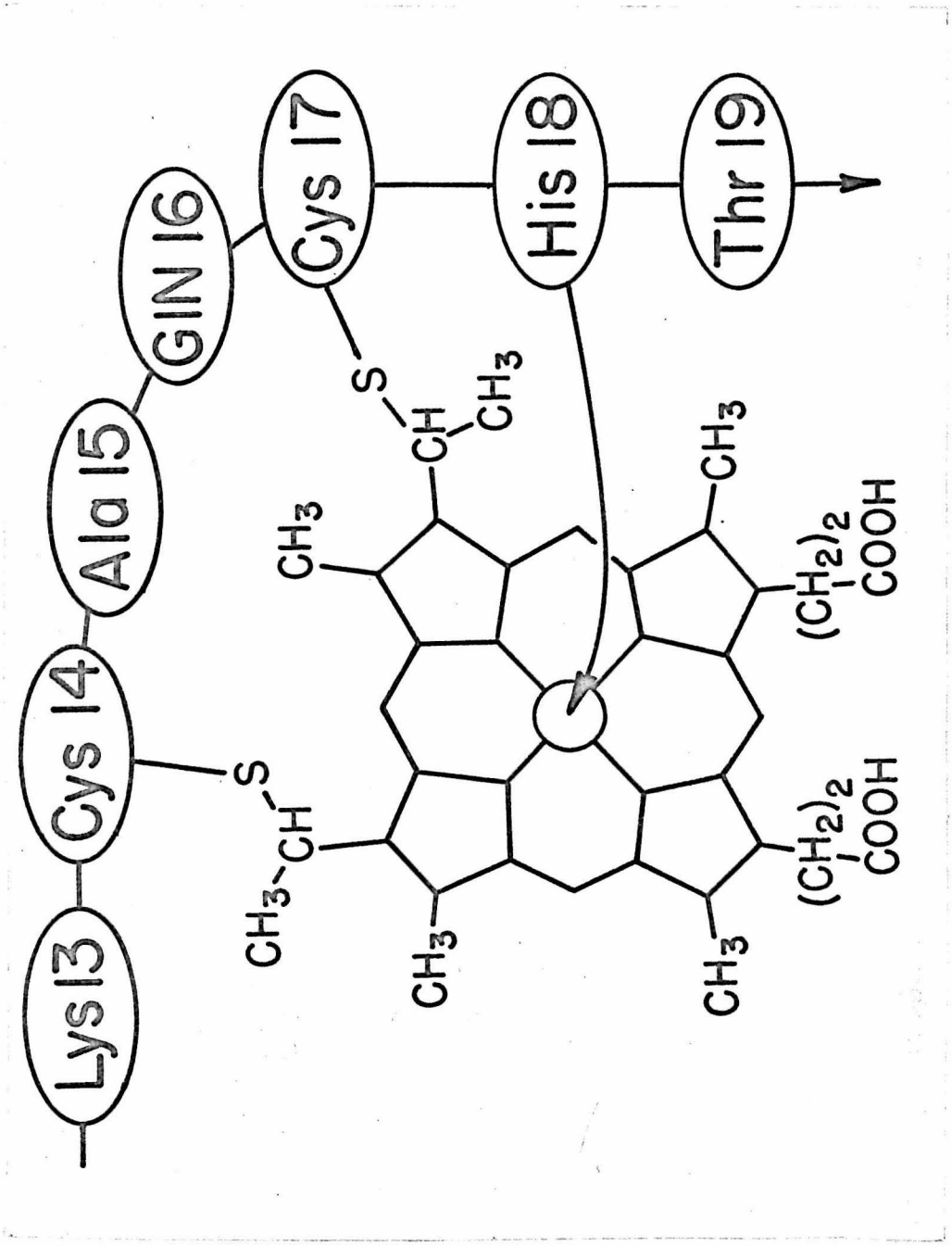
2) Wheat germ: E. Smith, private communication.



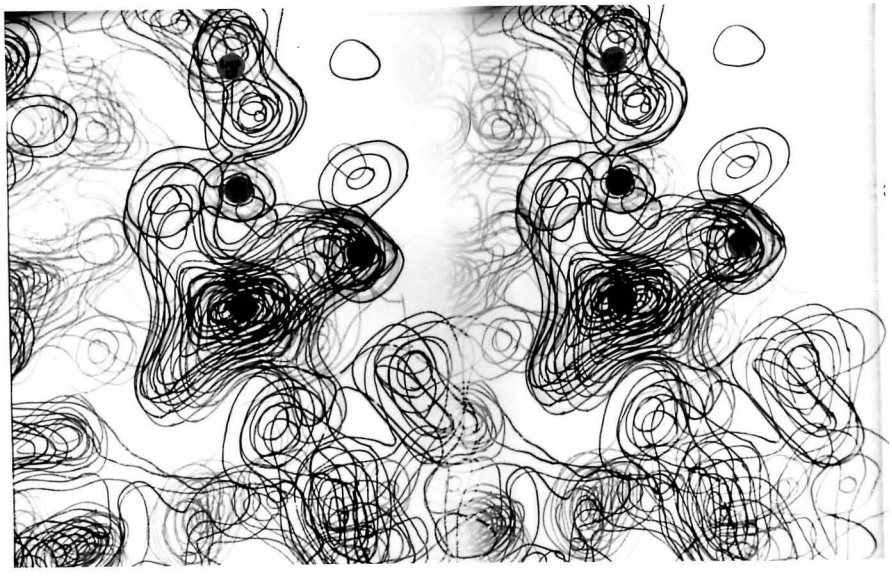
1.



3.

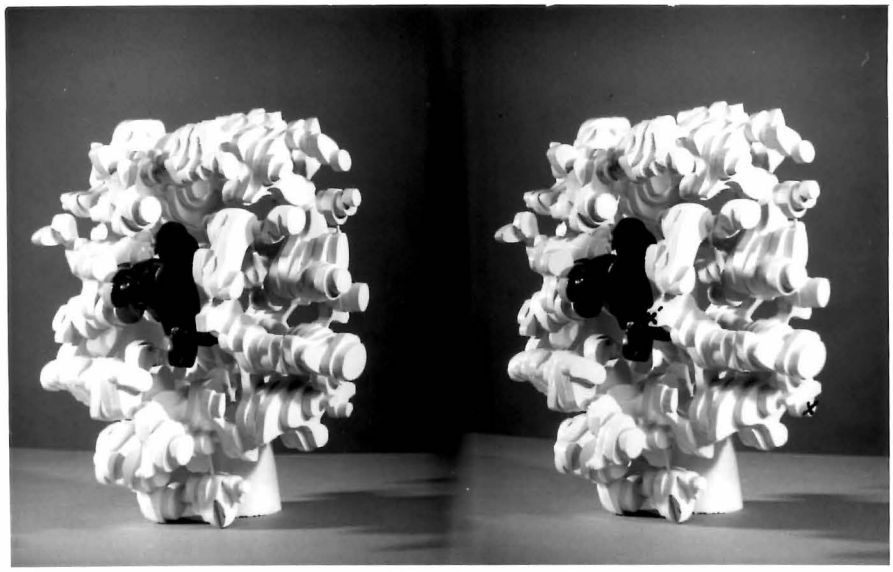

Lys 13
Cys 14
Ala 15
GIN 16
Cys 17
His 18
Thr 19
S
CH
CH3
CH3
CH3
CH3
CH3
CH3
CH3
(CH2)2
COOH
(CH2)2
COOH

2.

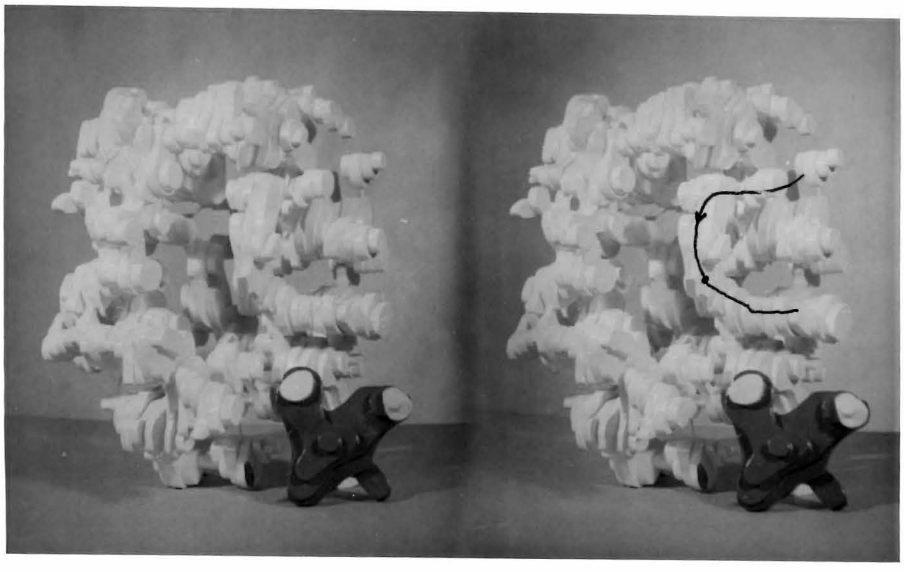


4.

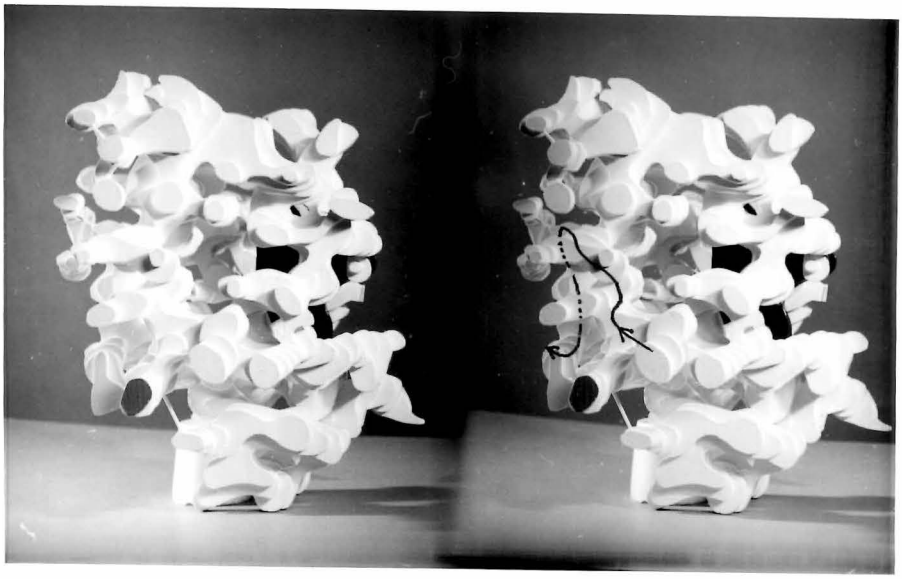
5.



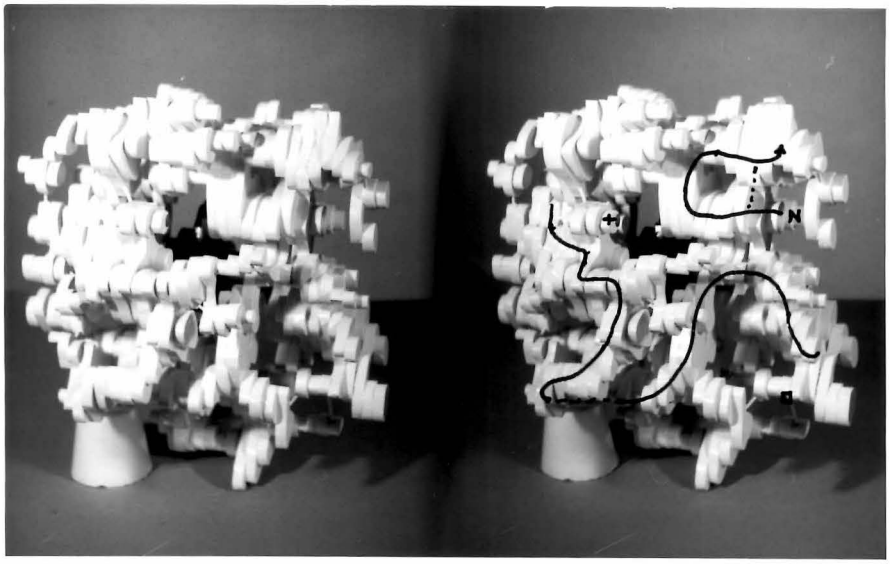
6.



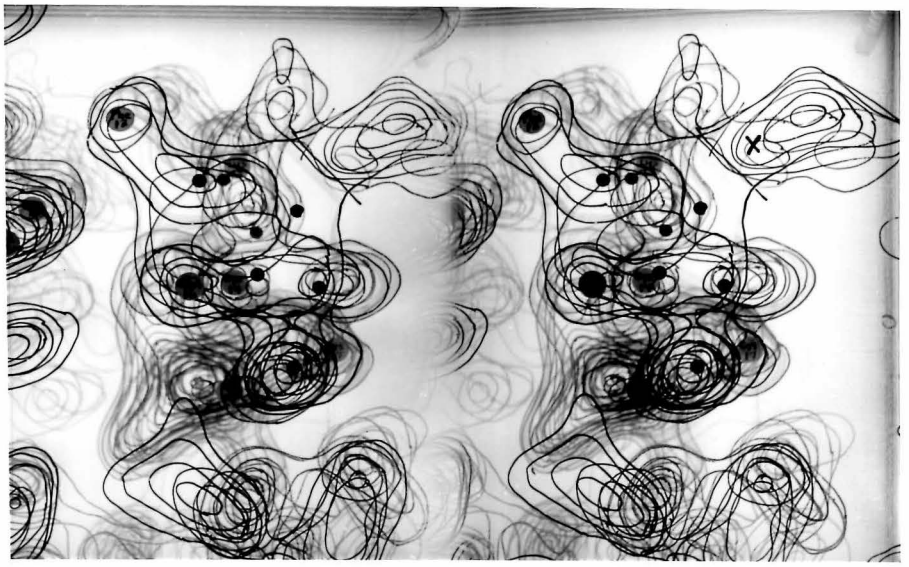
7.



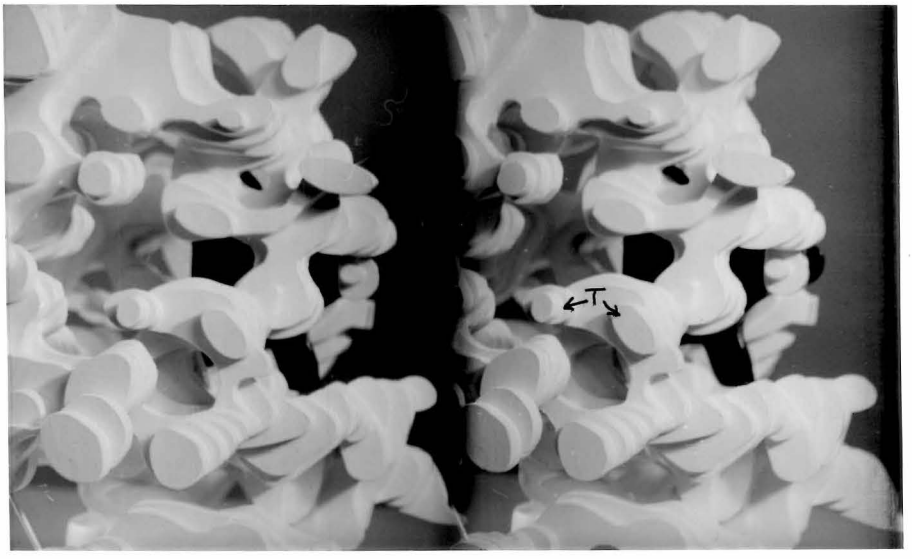
8.



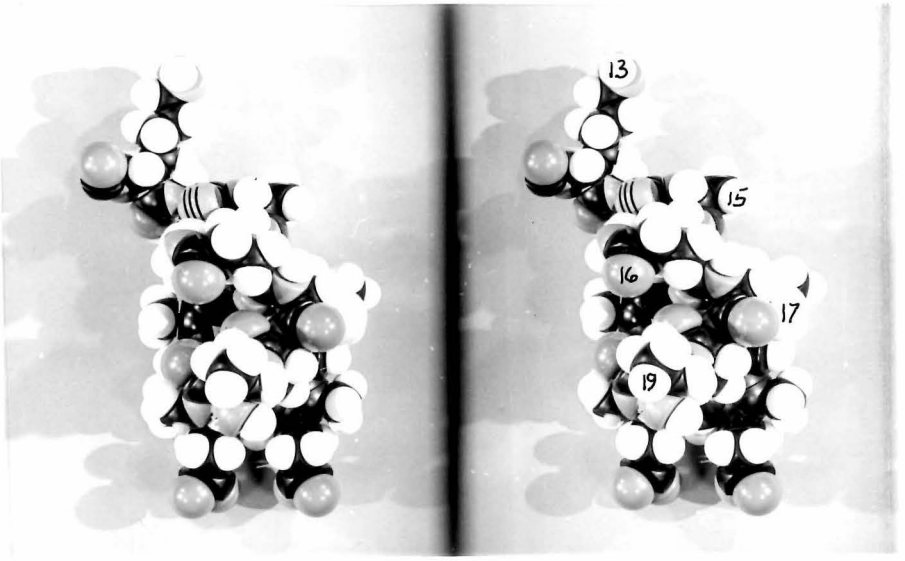

9.



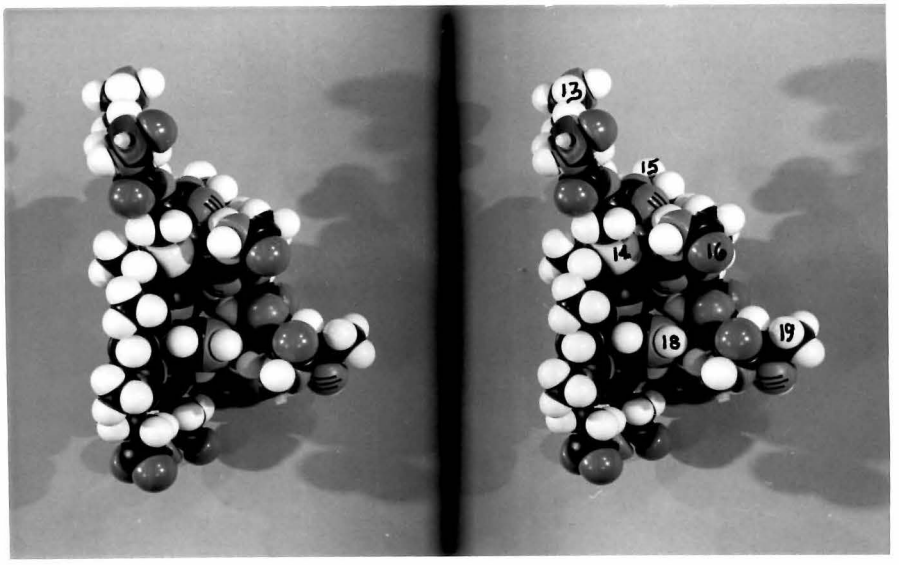

10.



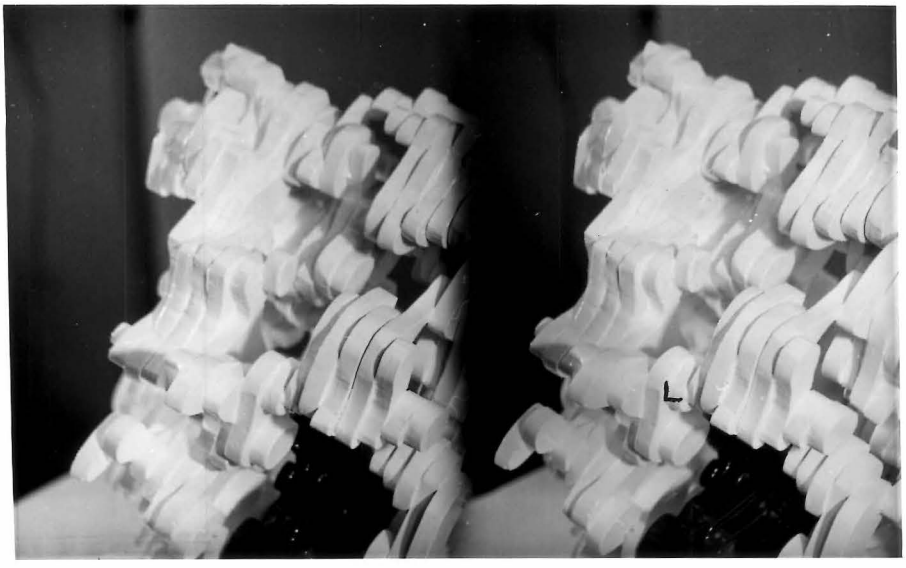

11.



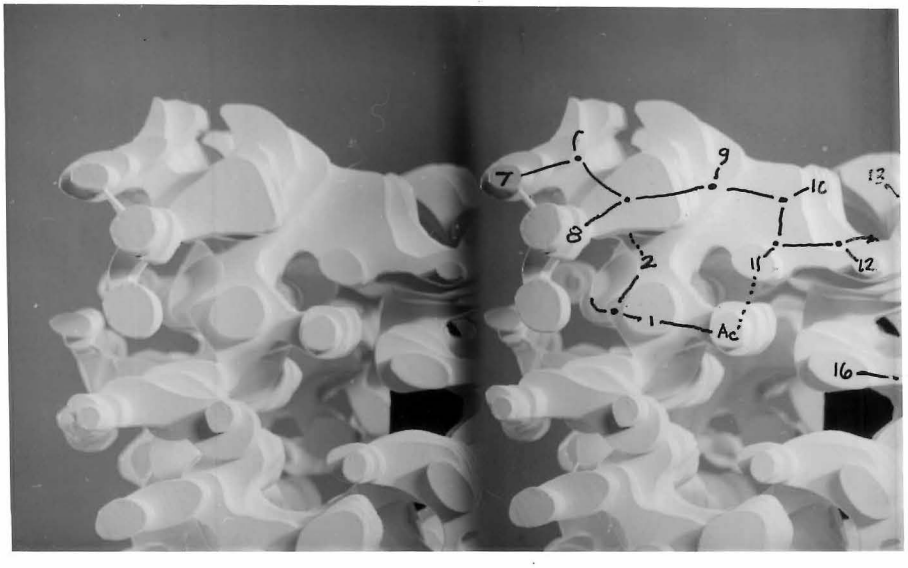

12.



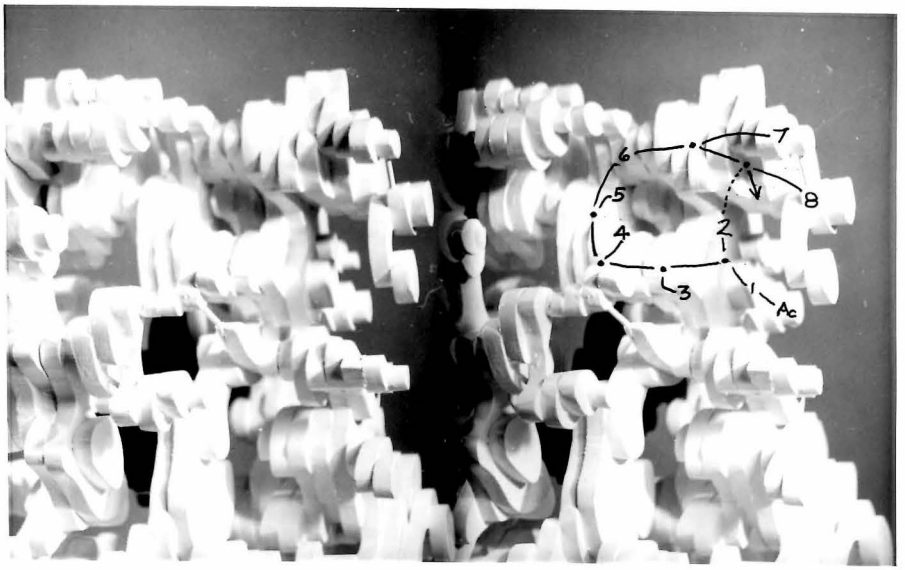

13.



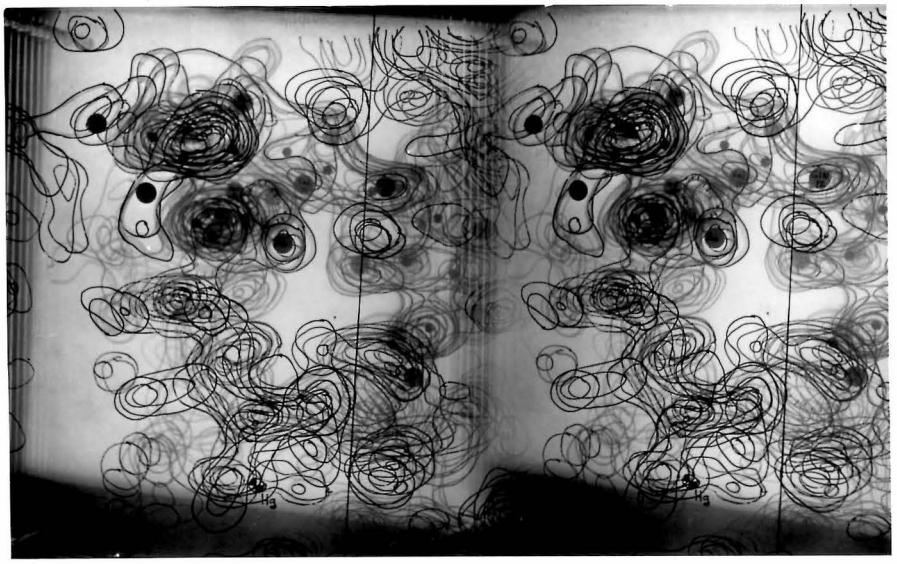

14.



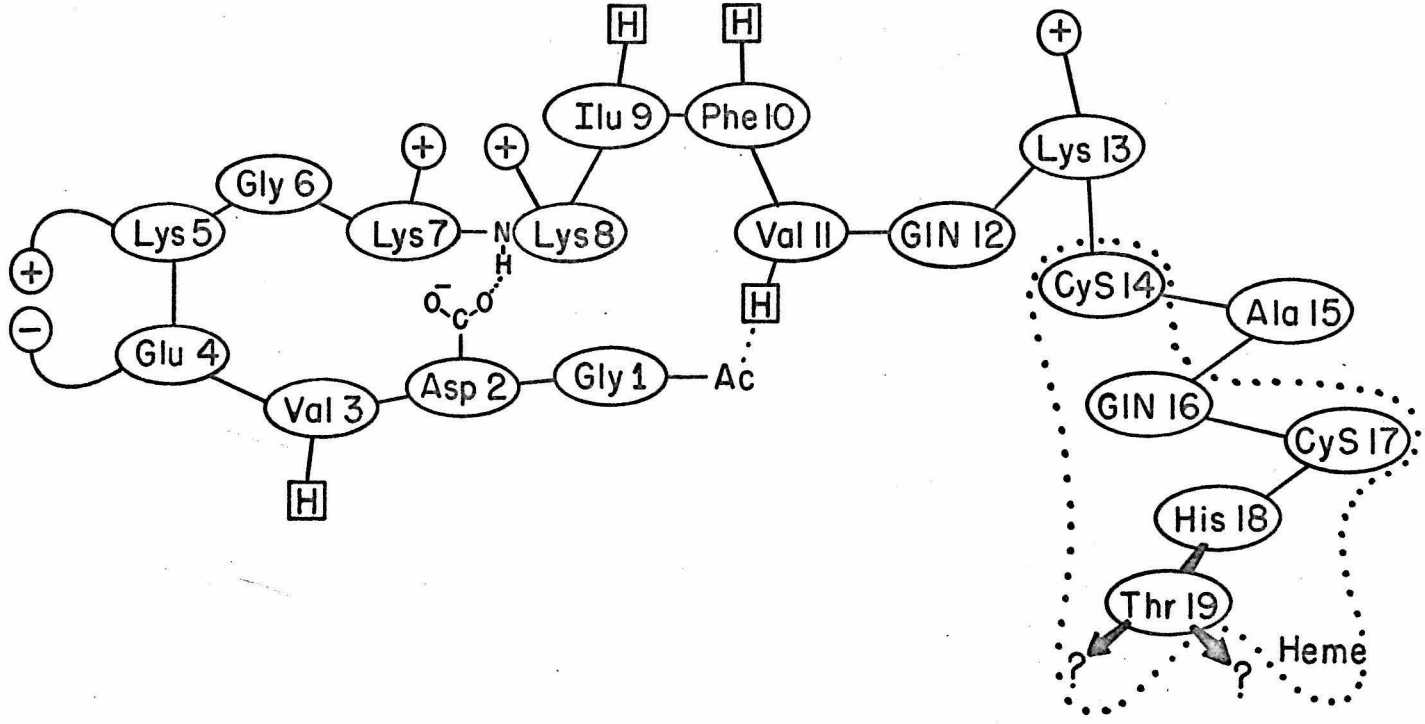

15.



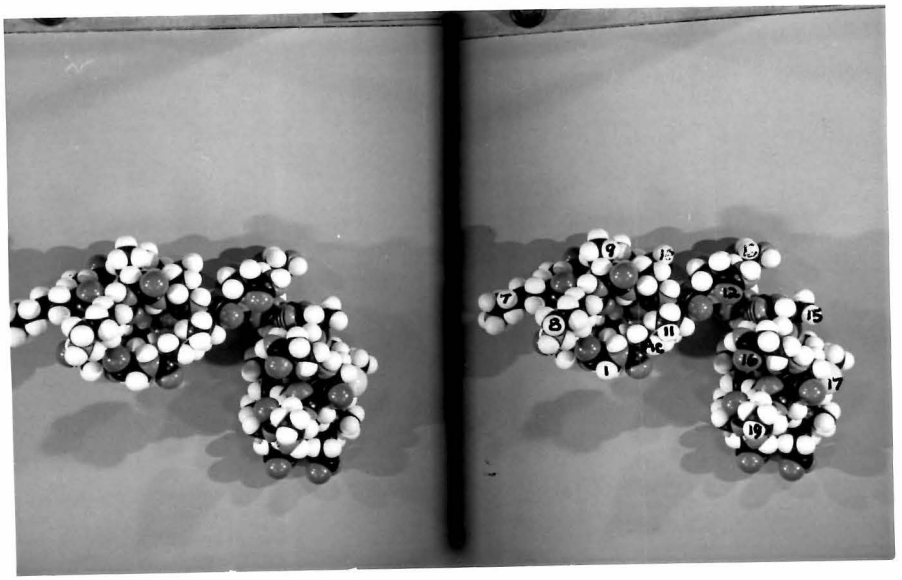

16.



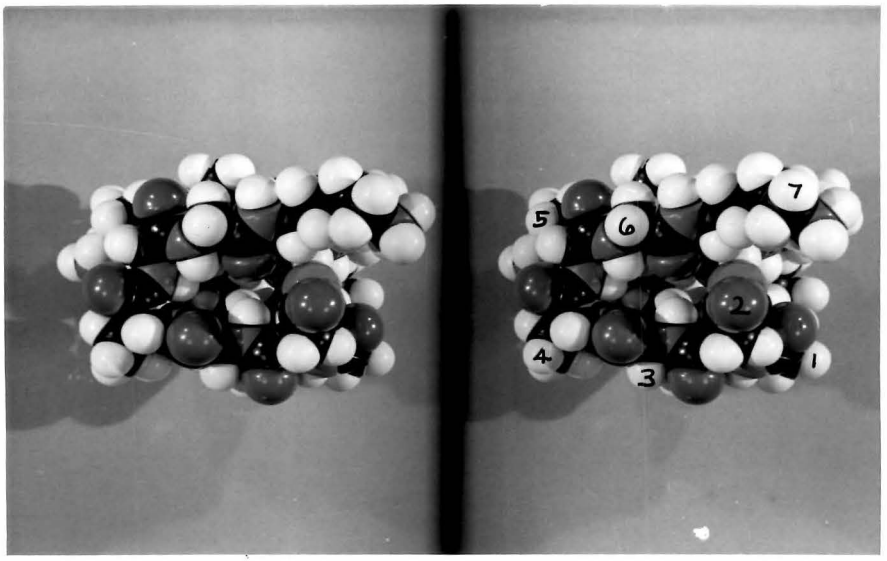
18


H
H
Ilu 9
Phe 10
Lys 13
Gly 6
Lys 5
Lys 7
N
H
Lys 8
Val 11
GIN 12
CyS 14
Ala 15
Glu 4
Asp 2
Gly 1
Ac
GIN 16
Val 3
CyS 17
His 18
Thr 19
Heme

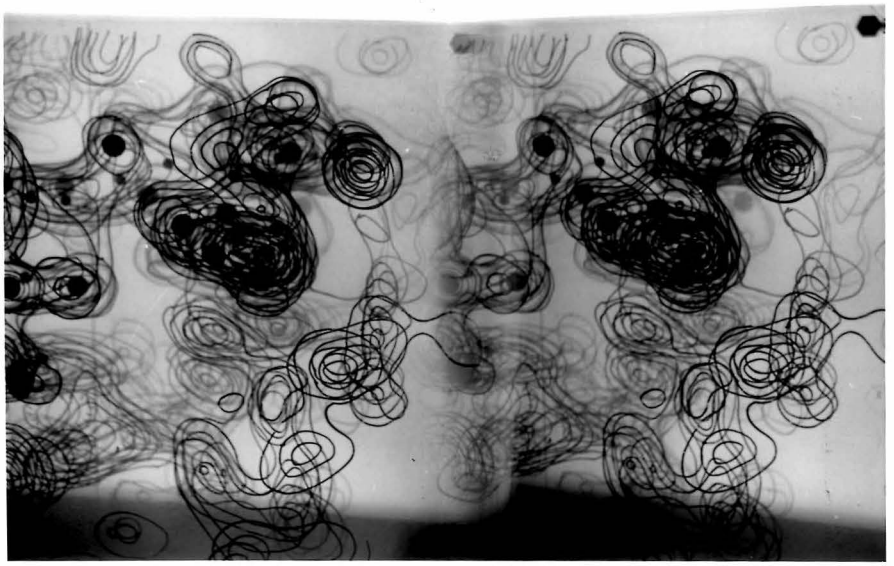
17.



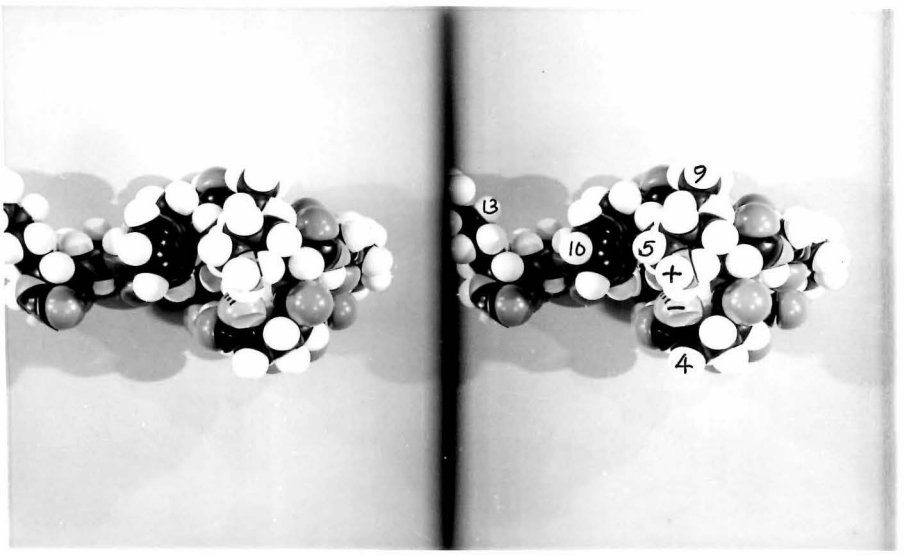

19.



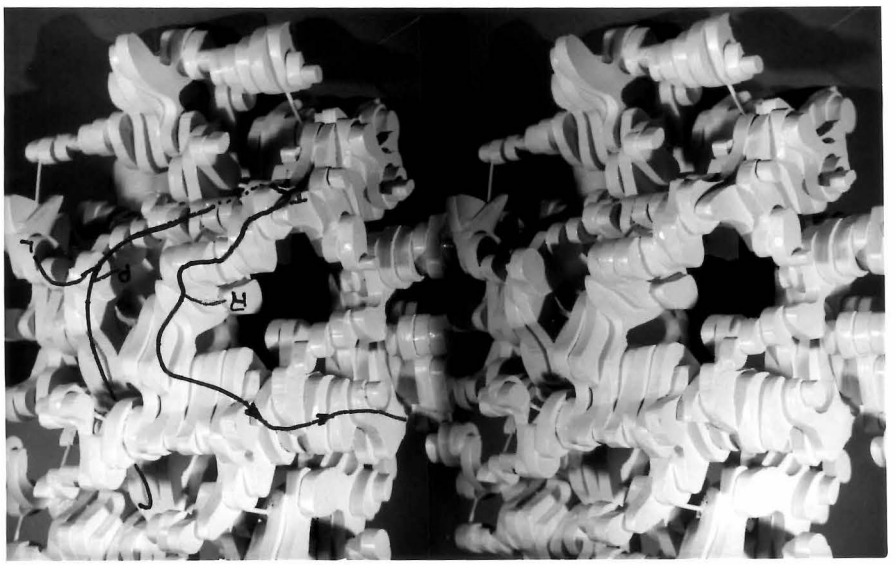

20.



21.




22.




23.

# Refinement of Protein Phases with the Karle-Hauptman Tangent Formula

JON E. WEINZIERL, DAVID EISENBERG AND RICHARD E. DICKERSON

Gates and Crellin Laboratories of Chemistry, †

California Institute of Technology, Pasadena, California, U.S.A.

† Contribution No. 3656 from the Gates and Crellin Laboratories of Chemistry.

Running title:

"Protein Phase Refinement with the Tangent Formula"

## Abstract

The Karle-Hauptman tangent formula has been tested as a means of refining phases previously determined from single and multiple isomorphous replacement. A test of the tangent formula, using exact phases of a model protein as a starting point, showed that the formula reproduces phase angles in good agreement with the true phases, even at only 4 Å resolution. When errors were introduced into the model phases, the tangent method refined back towards the true phases. This procedure was used to refine the phases of cytochrome c obtained from a double isomorphous analysis at 4 Å resolution, and some improvement in the electron density at the heme group was apparent.

Trials with single isomorphous replacement phases for a hypothetical derivative of the model protein showed that the tangent formula can be used to resolve the single-derivative ambiguity. Applying this method to single-derivative phases for cytochrome c, an electron density map was obtained which appears to be only slightly inferior to the two-derivative map.

## Introduction

The multiple isomorphous replacement (MIR) method of phase analysis, which made protein structure analysis possible, is still the standard method of finding protein phases. The trial and error method is manifestly impossible with molecules of this complexity, and Fourier refinement methods are useful only in the later high resolution stages. Patterson methods have proven useful only for finding heavy atom positions. The heavy atoms which occur naturally or can be added to protein molecules are not sufficiently heavy to dominate phasing. The anomalous scattering effect is smaller yet, and can be used only as an auxiliary to isomorphous replacement. The only significant variant to the MIR method has been the single isomorphous replacement (SIR) method with its inherent phase ambiguity broken by the use of anomalous scattering data (Kartha, 1961; Blow and Rossmann, 1961). Kartha showed that the SIR method alone is similar to vector superposition using the heavy atom positions, and that if the heavy atom constellation itself does not possess a center of symmetry, then the SIR map is built up from a number of images of the protein molecule in proper register, plus an equal number of images of the enantiomorph which add in a nonsystematic manner to give a high background noise level. Blow and Rossmann demonstrated that the map of hemoglobin produced by SIR and anomalous scattering, although considerably degraded from the seven compound MIR map, was still correct in gross features.

One class of phasing methods which has seen relatively little use with proteins is statistical phasing, of which the most powerful methods seem to

be those developed by Karle, Karle and Hauptman (Karle, 1964). In the two decades since these methods appeared, they have shown their worth with increasingly complex structures, first centrosymmetric and then non-centrosymmetric. But proteins have been relatively unexplored territory. C. L. Coulter (1965) has carried out model compound experiments along lines similar to some of those in this paper, and D. C. Phillips has carried out some unpublished experiments with lysozyme. The feeling has been prevalent that Karle-Hauptman methods could not produce protein phases ab initio, and that in any case they would not be applicable to data short of atomic resolution, or the vicinity of 2-1.5 Å. We have accepted the first restriction, but have found the second not to be true under proper conditions.

We have attempted to use the Karle-Hauptman tangent formula (equation 4.5 of Karle and Karle, 1966, hereafter referred to as K and K) in two ways, first as a means of refining phases produced by the MIR method, and then as a means of choosing between the two possible solutions of the SIR method. In both cases, the procedure has first been checked on a dummy protein ("modelglobin") having structure factors and phases calculated from the backbone chain of myoglobin out to and including β-carbon atoms, and then applied to real data for horse heart cytochrome c. The data used extended out to 4 Å and to 3.7 Å, respectively.

## Test of the tangent formula with model protein phases

The principles of Karle-Hauptman phase refinement are given in Karle (1964) and in K and K, and only those equations which were of direct

use will be given here. The tangent formula is:

$$\tan \phi_h \simeq \frac{\sum_k |\underline{E}_k \underline{E}_{h-k}| \sin (\phi_k + \phi_{h-k})}{\sum_k |\underline{E}_k \underline{E}_{h-k}| \cos (\phi_k + \phi_{h-k})} \qquad (1)$$

where h and k each represent Miller index triples. The $\underline{E}$'s in equation 1 are normalized structure factor magnitudes as given by:

$$|\underline{E}_h|^2_{obs} = \frac{|\underline{F}_h|^2}{\epsilon \sum_j f_j^2 (h)} \qquad (2)$$

$|\underline{F}_h|$ is the observed structure factor magnitude of reflection h, $f_j$ is the atomic scattering factor for the $j^{th}$ atom in the unit cell and $\epsilon$ is a number which corrects mainly for the space group ~~extinctions~~ symmetry. Equation 1 provides a way of finding a new, refined phase for each reflection in terms of the entire set of phases which were initially assumed on the right hand side of the equation.

Before setting out to refine sets of approximate phase angles, the accuracy of the tangent formula itself at low resolution was tested with the exact phases of modelglobin. Structure factors and phases for the 1239 reflections within the 4 Å limit were calculated using the polypeptide backbone atoms, β-carbons, iron and heme group of myoglobin in its true unit cell. We are indebted to Dr. John C. Kendrew for the myoglobin coordinates.

These phases and normalized structure factors will be designated by $\phi_{obs}$ and $\underline{E}_{obs}$.

The phase angles of the 410 reflections of modelglobin with $\underline{E}_{obs}$ greater than 1.0 were calculated from equation 1, using $\phi_{obs}$ and $\underline{E}_{obs}$ of all 1239 reflections on the right hand side. The results are shown in Fig. 1. The mean error in calculated phase angle is only about 6° for the fifty reflections with largest $\underline{E}_{obs}$, but the error in phase increases as $\underline{E}_{obs}$ decreases. The tangent formula begins to fail at significantly higher $\underline{E}_{obs}$ values if the data are cut off at 5 Å. This suggests that resolution of the data is a problem, and that the tangent formula will be more useful at 2 Å or higher resolution. Even at low resolution, however, the tangent formula can be used to improve the phases of the strongest reflections; these are often poorly determined by the MIR method because the ratio of $|f_H|$ to $|F_P|$ is small. At low resolution, the Karle-Hauptman and isomorphous replacement methods are complementary.

Several functions were tried as possible indices of the correctness of phases determined by the tangent formula, but none of these proved entirely satisfactory. One such function, called the "unnormalized calculated $\underline{E}$", was defined by analogy with the tangent formula:

$$\left| \underline{E}_h \right|_c = \left| \sum_k \left| \underline{E}_k \underline{E}_{h-k} \right| \exp \{i (\phi_k + \phi_{h-k})\} \right| \qquad (3)$$

These $\underline{E}_c$'s asymptotically approached constant values as the number of terms in the tangent formula increased (Fig. 2). For modelglobin, the mean value of $\underline{E}_c / \underline{E}_{obs}$ for the 410 reflections with the largest $\underline{E}_{obs}$ was found to be

120, but values of this ratio for individual reflections ranged from 10 to 290. No correlation could be found between this ratio and the error in the corresponding phase from the tangent formula, except that when the ratio was less than 40, the phase angle was usually greatly in error. For this reason the Karle and Karle R index (equation 5.2 of K and K) would not seem to be useful here as a criterion of phase quality at low resolution.

The effect of the number of terms used in the tangent formula upon the calculated phase angle and $\underline{E}_c$ is shown in Fig. 2 for a centric and an acentric reflection. The terms were added in order of decreasing $\underline{E}_k$. These two reflections are substantially determined by the top 300 terms. In general, however, at least 700 terms were needed to calculate phases of reflections with $\underline{E}_{obs}$ values of 1.0 or greater.

Another possible measure of phase correctness is based on the phase probability function defined by Karle and Karle (1966):

$$P_{(\phi'_h)} = [2\pi I_0(\alpha)]^{-1} \exp\{\alpha \cos(\phi_h' - \phi_h\} \qquad (4)$$

This expression gives the probability of correctness of a chosen phase angle $\phi_h'$ as a function of the coefficient $\alpha$ and the distance of $\phi_h'$ from the phase calculated from the tangent formula, $\phi_h$. The sharpness of the probability distribution increases as $\alpha$ rises. The coefficient $\alpha$ is proportional to the product of observed and calculated $\underline{E}_h$ values:

$$\alpha = \frac{2\sigma_3}{\sigma_2^{3/2}} |\underline{E}_h|_c |\underline{E}_h|_{obs} \qquad (5)$$

where the sigmas are defined in terms of the atomic numbers, $Z_j$, as:

$$\sigma_x = \sum_j Z_j^x \qquad (6)$$

A plot of $\alpha$ against the error in determining a phase angle is shown in Fig. 3. Except for very large values of $\alpha$, the correlation between $\alpha$ and the correctness of a phase is quite weak, and it thus appears that $\alpha$ is not much better than $\underline{E}_c / \underline{E}_{obs}$ as a measure of phase quality. The variance of $\phi_h$, suggested as an index of phase correctness by Karle and Karle (1966), could be calculated easily from $\alpha$ and equation 3.33 of K and K, but because of the lack of correlation shown in Fig. 3, this was not done.

In the refinements of cytochrome c described below, the appearance of the heme group provided the best quality criterion. The progress of refinement was monitored by calculating a Fourier map after every 100 or so phases had been refined with the tangent formula. The refinement of modelglobin was judged by comparing refined phases with exact phases.

## Refinement trials with modelglobin

Refinement with the tangent formula was always carried out in descending order of $\underline{E}_{obs}$ values, and two modes of refinement were tried. In "block cycling", a specified number of phases were recalculated, using the starting phases on the right hand side. All of these phases were then substituted together for the old ones before a second cycle was calculated. In "accelerated cycling", a newly calculated phase was immediately substituted for its initial value, where it would then contribute to all reflections of lesser $\underline{E}$ value.

As a preliminary experiment in refinement, the fifty acentric reflections of largest $\underline{E}_{obs}$ values were subjected to repeated block and accelerated cycling, with the results shown in Fig. 4. The change in phase from one cycle to another is diminishing, and the phases are asymtotically approaching a set which is self-consistent but which differs from the correct set by about twelve degrees. It may be that this difference between self-consistency and correctness is a consequence of the limited amount of data used, and that at high resolution this anomaly would not arise.

Two trials were made in which random errors were introduced into the calculated phase set, and the tangent formula was used to refine back again. In the first trial, errors in acentric reflections ranged from 0° to 90° with a mean of 32°, and the signs of two percent of the centric reflections were reversed. The distribution of errors introduced into the acentric reflections is shown in Fig. 5. The refinement behavior of the 50 largest acentric reflections and then of the 147 largest acentric reflections when subjected to block and accelerated cycling is shown in Fig. 6. The mean phase error drops dramatically in the first cycle or two, levels off, and then begins to rise again with block cycling. Accelerated cycling reduces the mean error more rapidly, but also brings on the subsequent "blowup" of phase angles sooner.

Similar behavior was found in a second set of trials, in which the phases of the most intense 100 reflections were put in error by an average of 56° and the other phases were left as in the 32° trial. Accelerated cycling of the top 72 reflections reduced the mean error from 56° to 31°, then to 27.9° and 28.0° in two more cycles. With the greater initial error in phases, the onset of the drift toward the self-consistent set is delayed compared to the previous run.

The origin of this phase blowup upon continued cycling is not known. One possible explanation might be the one alluded to earlier; with limited data the self-consistent set may differ from the correct one. The rise in phase error after several cycles may be a drive towards self-consistency. This trend may also be present at the beginning, but be masked by improvement in phase angles which were initially badly in error. Only higher resolution data will resolve these questions.

The conclusion from these modelglobin trials is that refinement of MIR phases at low resolution should be carried out for two or three cycles for the strongest reflections and then halted before phase deterioration begins. If this is done, then a certain amount of phase improvement will result.

## MIR phase refinement with cytochrome c

Unlike many proteins, cytochrome c has a prosthetic group, the heme, which is readily recognizable even at 4 Å resolution. The appearance of the heme group was used as a criterion of improvement during tangent formula refinement.

Two cytochrome derivatives were available, Pt and Hg, each with a different single binding site to the protein molecule. Pt was well-substituted, with a mean change in structure factor by the heavy atom of 29% in the centric hk0 zone of space group $P4_1$. Hg was weaker, with a mean change of only 15%. Further details about data, derivatives and MIR phase analysis will be found in Dickerson, Kopka, Borders, Varnum, Weinzierl and Margoliash (1967) and Dickerson, Kopka, Weinzierl, Eisenberg and Margoliash (1967).

Preparation of normalized structure factors, $\underline{E}_{obs}$, presented a problem at low resolution. With the peaks in the $\underline{F}$ curve at about 10 Å and

5 Å, Wilson plot scaling was impossible. Instead, a temperature factor of exp $(-0.475S^2)$ was assumed to be present by analogy with Kendrew's findings for myoglobin, and was removed by multiplying the data by exp $(0.475S^2)$. $(S = 2 \sin \theta)$. The data were then scaled so that $\langle \underline{E}_{obs}^2 \rangle = 1.0$ for the region from 8 Å to 4 Å. The inner reflections out to a resolution of 13 Å were removed because of their sensitivity to the salt concentration of the crystallizing medium.

Refinement of the MIR phases by accelerated cycling was tried first. After every 100 reflections, a Fourier map of the heme region was calculated. As the refinement proceeded, the iron atom gained in peak intensity in successive Fourier maps and the detail of the surrounding heme was steadily degraded. Block cycling was then tried, with much better results. The strongest 197 reflections (out of 1439 at 3.7 Å resolution) were given one cycle of block refinement. These 197 reflections were used with the unchanged phases of the 1242 reflections of lesser $\underline{E}_{obs}$ to calculate a much more encouraging map around the heme region. Refinement of the next 50 reflections did not appreciably change the map, nor did a second cycle of the top 100 reflections.

It appeared to make little difference whether or not the MIR figures of merit were used to weight the right hand side of the tangent formula at this point, and in the final calculations they were not used. The SIR work to follow, however, suggested that weighting with figures of merit is the proper thing to do. The refined 197 reflections were given a new figure of merit for the purpose of calculating Fourier maps, defined as:

$$m' = \frac{1}{2} + \frac{1}{2} \cos \pi \left(m - \cos \left(\frac{|\Delta\phi|}{2}\right)\right) \qquad (7)$$

The justification for this rather arbitrary function was that the weight should range between 0 and 1, and that a reflection should be given a large weight if it had had a large figure of merit and had not been altered appreciably by tangent refinement, or if it had had a small figure of merit and had been changed considerably. Non-refined reflections retained their old figures of merit. Unweighted maps calculated as controls showed the same general trends as weighted maps.

Of the several maps which were calculated, two are particularly informative: the original unrefined map weighted with figures of merit, and the refined map weighted with new figures of merit. These two maps are shown in Figs. 7 and 8, and 10 and 11, and a theoretical heme at several resolutions is in Fig. 9. Points of identification for the heme maps include the high density at the iron atom, the symmetrical extension of the two propionic acid side chains, the asymmetrical arrangement of the two thioether links to the protein, the polypeptide chain sequence: -Cys14-Ala15-Gln16-Cys17-His18- running from the upper thioether link to the one at the right, and from there to the fifth iron coordination position on the near side of the heme, and the curved polypeptide chain running behind the heme and bearing the side group which makes the sixth coordination with the iron.

In plan view, the heme is not drastically altered other than in having more density at the iron and being generally more skeletal in appearance. But the view from the upper edge (Figs. 10 and 11) shows the differences more clearly. The density and connectedness of the polypeptide chain sequence between thioether links is greatly increased in the refined map. The two maps were put on the same scale by Wilson plots so that relative densities

would be comparable. As a result of refinement, the peak iron density rose from 500 to 610 on an arbitrary scale, and the two sulfur sites became more consistent, that of Cys14 rising from 196 to 282 and that of Cys17 falling from 338 to 288. The density at the sixth coordination site, which has been thought on chemical grounds to be methionine, rose from 130 to 205. The only undesirable feature of the refined map is that the extended chain with the sixth ligand was weakened.

## Resolution of the SIR phase ambiguity with the tangent formula

The results of the refinement of MIR phases were encouraging but not conclusive. It appeared that some improvement in the appearance of the heme in cytochrome had been produced. It was also apparent that working at low resolution meant working just this side of disaster, and that the method would be considerably more useful at high resolution.

The Karle-Hauptman method could be more powerful at low resolution, paradoxically, if it were asked to do less. Instead of being used to calculate a new phase angle, could it usefully be employed only to indicate which of the two possible phases in a SIR analysis was the correct one?

The principles of SIR phasing are illustrated in Fig. 12. The use of the centroid vector S is equivalent to choosing half of each of the phase alternatives. The map will then be half correct density and half incoherent noise. Any means of choosing phase alternates would greatly improve the map. If the bimodal probability curve around $\phi$ were perfectly sharp, then the centroid figure of merit would be equal to $|\cos \delta|$. A traditional heavy atom

Fourier map of the protein would be built up from coefficients: $F_P \exp(i\phi_H)$. The SIR map coefficients can be thought of either as: $m F_P \exp(i\phi_H)$, or as: $|m| F_P \exp(i\phi_S)$, where $m = \cos\delta$ and is positive if $\delta$ is less than 90° or negative if it is greater. In this way of thinking, the function of the addition of parent compound data to an isomorphous heavy atom derivative is to provide the proper weighting factors for the heavy atom Fourier map.

The tangent formula can be used to calculate phase angles, $\phi_c$ using SIR phases weighted with figures of merit on the right hand side of the equation. If it is assumed that the SIR method is accurate but ambiguous and the Karle-Hauptman method is definite but inaccurate, then the best strategy is to choose the SIR phase which lies nearest to $\phi_c$.

The problem of weighting factors to replace the SIR figures of merit remains. But if $\delta$ is small, then it matters little whether the Karle-Hauptman phase is indicative or not. On the other hand, if $\phi_c$ falls very near to one phase choice and $\eta$ is small, then it is irrelevant how large the SIR phase ambiguity initially was. The new figure of merit in the cytochrome experiments was therefore chosen to be either $\cos\delta$ or $\cos\eta$, whichever was greater.

## Test of the phase choice method with the model protein

In the tests of the method with modelglobin, SIR phases were generated for a hypothetical two-site derivative having mercuric ions at (0.275, 0.250, 0.142) and (0.400, 0.412, 0.485) and the symmetry related positions of the

myoglobin unit cell.* These are actually the PCMBS and Au positions from sperm whale myoglobin (Bodo, Dintzis, Kendrew, and Wyckoff 1959). The SIR phases and cos δ weights were used in the tangent formula to calculate a new set of phases, $\phi_c$, in order of decreasing $\underline{E}_{obs}$. Each $\phi_c$ was used immediately upon calculation to choose between the SIR phase alternatives, and this chosen phase was substituted for the initial SIR value in a variation of accelerated cycling.

Accelerated cycling was used in spite of our experience with MIR refinement because of the way in which the tangent formula phases were used. It was hoped that the additional factor, that the phases actually used were from isomorphous replacement, would keep the refinement out of the self-consistency trap and would make phasing meaningful down to much lower $\underline{E}_{obs}$ values than before.

Fig. 13 shows the results of accelerated SIR refinement of the 700 reflections of largest $\underline{E}_{obs}$ in the 4 å modelglobin data set. Although they include only half of the reflections, they include about 85% of the total $\underline{E}_{obs}$ sum. The angle between a SIR phase and one of its possible phase choices can vary from 0° to 90°, so the SIR set before refinement will be something

---

* It should be noted that this method of SIR phase choice cannot be used if the heavy atom cluster is centrosymmetric, for then there is no way of driving $\phi_c$ away from the real axis. This situation arises, for example, in space group $P2_1$ with a single site derivative, but not in $P4_1$. A method of circumventing this difficulty has been suggested by Karle (1966).

like an average of 45° in error. After one accelerated pass, this error was reduced to a mean of 14.5° for the top 600 reflections and 35° for the next 100. The second accelerated cycle dropped the mean error by another 3.2°.

Expectations that the SIR choice procedure would be valid for a much greater range of E than the MIR refinement seem to be amply borne out. The modelglobin trials, of course, tell nothing about the effect of experimental errors on the refinement process; this aspect is best studied with real data.

## Application of SIR/KH phasing to cytochrome c

Since the electron density in the heme region of cytochrome c was to be used as the quality criterion of tangent formula phases, several auxiliary maps of the heme region were first computed as controls. In Fig. 14 is shown the Pt SIR map, which should be compared with the unrefined MIR map of Fig. 7. The heme in Fig. 14 is clearly visible, but is badly distorted. The center of the iron (peak density 450 rather than 500) is no longer the highest point in the map, the heme is emaciated in appearance, the Cys connections are erratic and the propionic acid side chains at the bottom are both continuous with a mass of density below the heme. The extended chain on the far side of the heme is wiped out. The equivalent Hg SIR map reduces the heme to little more than a flattened blob, although the extended chain in back is visible. Fig. 15 shows a map produced not by incorporating Hg data into the centroid phase analysis as in Fig. 7, but by merely using the Hg phase intersections to choose between the two Pt possibilities. The heme is very little improved over the Pt SIR map except for better definition of the propionic acid groups. The extended chain is broken but is present.

The success of accelerated cycling in resolving the SIR ambiguity in modelglobin led us to apply the same method to the Pt SIR results for cytochrome c. Reflections 1-450 were given two cycles of refinement, then reflections 451-800 were given one cycle, and finally reflections 1-100 were given a third cycle. In computing the Fourier map, revised figures of merit of section 5 were used, and the unrefined reflections of low E and resolution lower than 13 Å were included with their SIR phases and figures of merit. The results, shown in Fig. 16, were disappointing. The heme was distorted, its upper left corner was nearly wiped out, and the density of the iron position was increased from 500 on an arbitrary scale in the MIR map of Fig. 7 to 750.

Block cycling of the top 800 reflections was then tried instead of accelerated cycling, and as in the MIR refinements, it was much more successful. The resulting map is shown in Fig. 17; in contrast to Fig. 16, it is far superior to the Pt SIR map and is almost as good as the complete MIR map. The electron density at the iron has increased only to 600, and the features around the periphery of the heme are clear and well-formed. A final pass of block cycling through the top 100 reflections made no further improvement in the heme, and retained a peak density of 600 at the iron.

One final use was made of the Karle-Hauptman method. The angle $\phi_c$ and the coefficient $\alpha$ were calculated from the tangent formula and then used in equation 4 with the pre-exponential Bessel function omitted to calculate the relative Karle-Hauptman probabilities of phase angle $\phi'$ around the phase circle. This probability curve and the normal Blow-Crick SIR phase probability curve were then multiplied to form the joint probability, and the centroid of

this probability function found in the normal way. After block refinement of the top 800 terms, $\phi_c$ and $\alpha$ were fed into a modified phase program and centroid phases and figures of merit were produced. It was hoped that this would be the best way to merge isomorphous and statistical phase information, but in fact the map produced was the unfavorable one shown in Fig. 18. The density at the iron increased to 950, there was a termination-of-series ripple around the iron, and the map was closer to Fig. 16 than to Fig. 17.

## Conclusions

These experiments show that the Karle-Hauptman tangent formula is useful in several ways in refining protein phases. Once phases have been obtained from a low resolution MIR study, those for the strongest reflections (which are usually among the least well-determined phases) can be improved by one or two block cycles with the tangent formula. Further block cycling or accelerated cycling tends to degrade the map, and to produce a self-consistent but wrong phase set which piles additional density on the strongest features of the map. This difficulty has also been encountered in analyses of smaller structures at atomic resolution. It may arise from the constraints imposed by the tangent formula itself, such as the assumptions of equal point atoms and nonnegative electron density.

The use of the tangent formula to refine high resolution protein data has not been explored yet, but a comparison of results at 5 Å and at 4 Å suggests that a much greater fraction of the $\underline{E}$'s can be refined at higher resolution. At high resolution, the coefficient $\alpha$ for a reflection may be

more strongly correlated with the correctness of its phase angle, and it may then be more appropriate to incorporate the Karle-Hauptman results directly into the phase program as outlined in the previous section.

The tangent formula is more effective in the SIR application because it is asked to choose between two isomorphous phases and not to predict a numerical value. Our trials with modelglobin show that under favorable conditions this method can break the SIR ambiguity for the majority of reflections, even at low resolution. In a situation where there is one excellent single-site derivative of good isomorphism out to high resolution, and a collection of other poorer multiple-site derivatives, the wisest strategy may be to throw out the poor derivatives, recollect parent and single-site data again and again to eliminate as much experimental error as possible, and then to use Karle-Hauptman methods to resolve the phase ambiguity.

## Acknowledgements

We should like to express our appreciation to Dr. Richard H. Stanford for the seminar which initially triggered our interest in Karle-Hauptman methods, and for discussions at all stages of the work. We should also like to thank Drs. J. Karle and I. Karle for their helpful comments and suggestions during the A. C. A. meeting in Tucson.

This work was performed under the auspices of National Science Foundation research grant GB-3053, the help of which is greatly appreciated. One of us (JEW) is also the holder of a National Institutes of Health Predoctoral Traineeship (GM-1262).

## References

BLOW, D.M. and ROSSMANN, M.G. (1961). Acta Cryst. 14, 1195.

BODO, G., DINTZIS, H.M., KENDREW, J.C. and WYCKOFF, H.W. (1959). Proc. Roy. Soc. A253, 70.

COULTER, C.L. (1965). J. Mol. Biol. 12, 292.

DICKERSON, R.E., KOPKA, M.L., BORDERS, C.L., Jr., VARNUM, J.C., WEINZIERL, J.E. and MARGOLIASH, E. (1967). J. Mol. Biol. 29, 77.

DICKERSON, R.E., KOPKA, M.L., WEINZIERL, J.E., EISENBERG, D. and MARGOLIASH, E. (1967). J. Biol. Chem. 242, 3015.

KARLE, J. (1964). In Advances in Structure Research by Diffraction Methods, Vol. I. R. Brill, ed., New York, Interscience (Wylie).

KARLE, J. (1966). Acta Cryst. 21, 273.

KARLE, J. and KARLE, I. (1966). Acta Cryst. 21, 849.

KARTHA, G. (1961). Acta Cryst. 14, 680.

## Figures

Figure 1. The error in calculating phase angles of a model protein from the tangent formula, as a function of the normalized structure factor magnitude, using correct phases in the right hand side of the tangent formula. Each point represents the average over ten reflections.

Figure 2. The dependence of $\underline{E}_c$ and the calculated phase angle $\phi_c$ upon the number of terms used in the tangent formula. Terms were added in order of decreasing $\underline{E}_k$.

a. Centric 002 reflection.

b. Acentric 221 reflection.

Figure 3. The dependence of the coefficient $\alpha$ upon $|\phi_c - \phi_{obs}|$ for the modelglobin reflections with $\underline{E}_{obs}$ greater than 1.0.

Figure 4. Refinement behavior of modelglobin phases using the tangent formula, starting with the correct phases. Note that the apparent point of self-consistency differs from the true phase set by a mean of about 12°.

Figure 5. Distribution of the random errors introduced into the model protein phases as a test of tangent formula refinement.

Figure 6. Refinement behavior of the modelglobin phases after introduction of the errors of Fig. 5. Note the initial rapid return of phases toward the true model set, followed by their slower subsequent divergence.

Figure 7. The heme region of cytochrome c. This map is the result of a centroid phase analysis using Pt and Hg derivatives, and is weighted with figures of merit. See text for interpretation.

Figure 8. Same heme region, after Karle-Hauptman refinement. Note the enhancement of the central iron atom.

Figure 9. The heme skeleton of cytochrome c, and its expected appearance at several resolutions. For each map, heme structure factors were calculated in the true cytochrome cell, the data were cut off at the desired resolution, and a Fourier map was calculated. The propionic acid side chains at the bottom were assumed here to be in their most extended conformation.

a. Heme skeleton
b. 6 Å resolution
c. 5 Å
d. 3.7 Å

Figure 10. The heme region of the map of Fig. 7, plotted in sections normal to z instead of to x, and viewed down from the top right of Fig. 7. The heme is seen nearly edge-on, with the Cys17 bridge coming out from the center and curving to the left. The Cys14 bridge rises in the back and curves forward and to the left through the peptide chain to join Cys17 and then to drop down to His18, which is coordinated to the heme iron from the left. The extended polypeptide chain which carries the sixth ligand curves upward past the heme plane on the right.

Figure 11. The map of Fig. 8 (after Karle-Hauptman refinement), seen as in Fig. 10. Note the enhancement of density in the regions of

the iron, the sulfurs of Cys14 and Cys17, and the sixth iron ligand on the right. Note also the strengthening of the polypeptide chain between Cys14 and Cys17, and the weakening of the extended chain on the right.

Figure 12. Single isomorphous replacement phase diagram. Protein, heavy atom and derivative structure factor vectors are labeled P, H and PH, respectively. The two possible phase angles are $\phi_1$ and $\phi_2$. The SIR centroid vector (in the limit of perfectly sharp probability peaks) is labeled S and the SIR phase is $\phi_S$. The phase calculated from the tangent formula is $\phi_c$. (All $\phi$ angles are measured from the real axis.) $\delta$ measures the half angle of the SIR phase ambiguity, and $\eta$ is the angle between $\phi_c$ and the nearer of the two SIR phase possibilities.

Figure 13. Refinement behavior of the modelglobin acentric SIR phases. The first accelerated cycle removes nearly 70% of the error contained in the acentric phases associated with the 600 largest $\underline{E}$'s; the second cycle makes a small additional improvement.

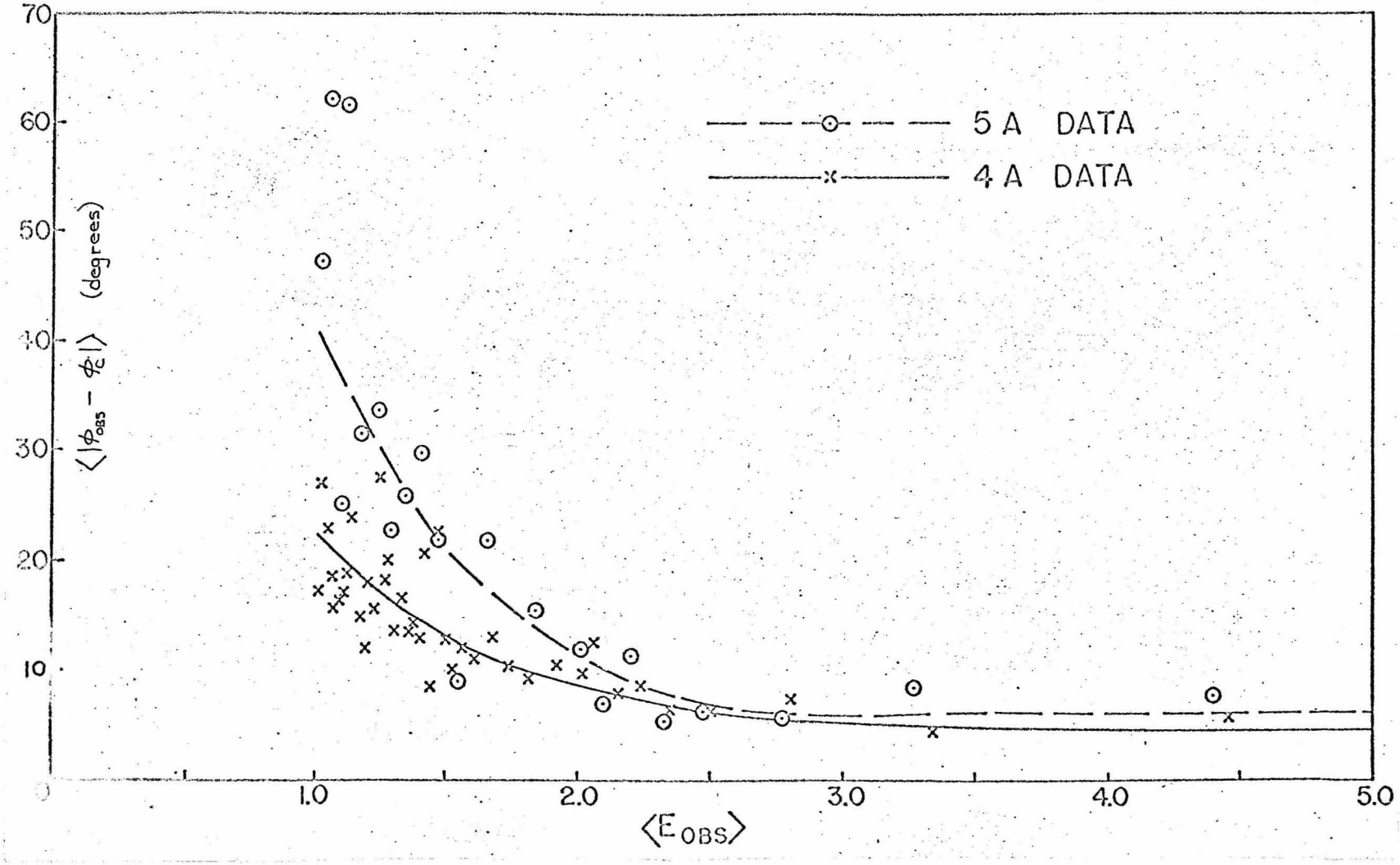
Figure 14. Heme region of cytochrome $\underline{c}$ using Pt SIR phases and figures of merit.

Figure 15. Heme region using the Hg phase possibilities to choose between the two possible Pt SIR phases.

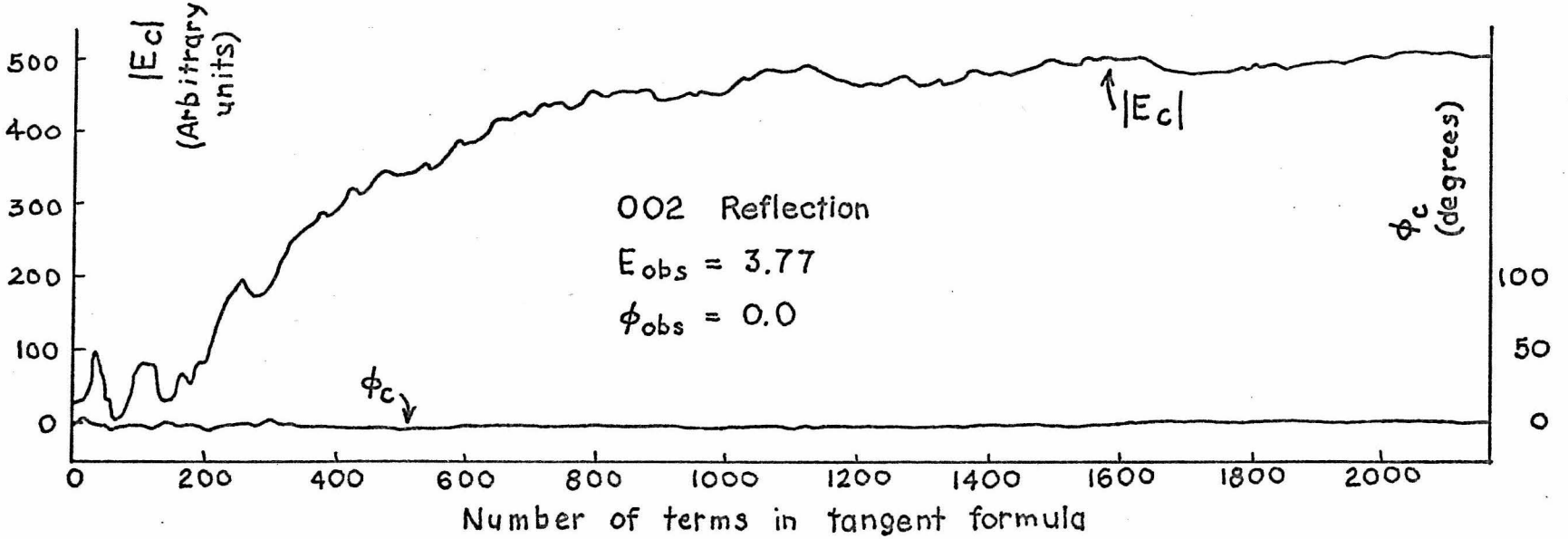
Figure 16. Pt SIR map after tangent formula refinement of the top 800 terms, with accelerated cycling.

Figure 17. Pt SIR map after tangent formula refinement of the top 800 terms, with block cycling.

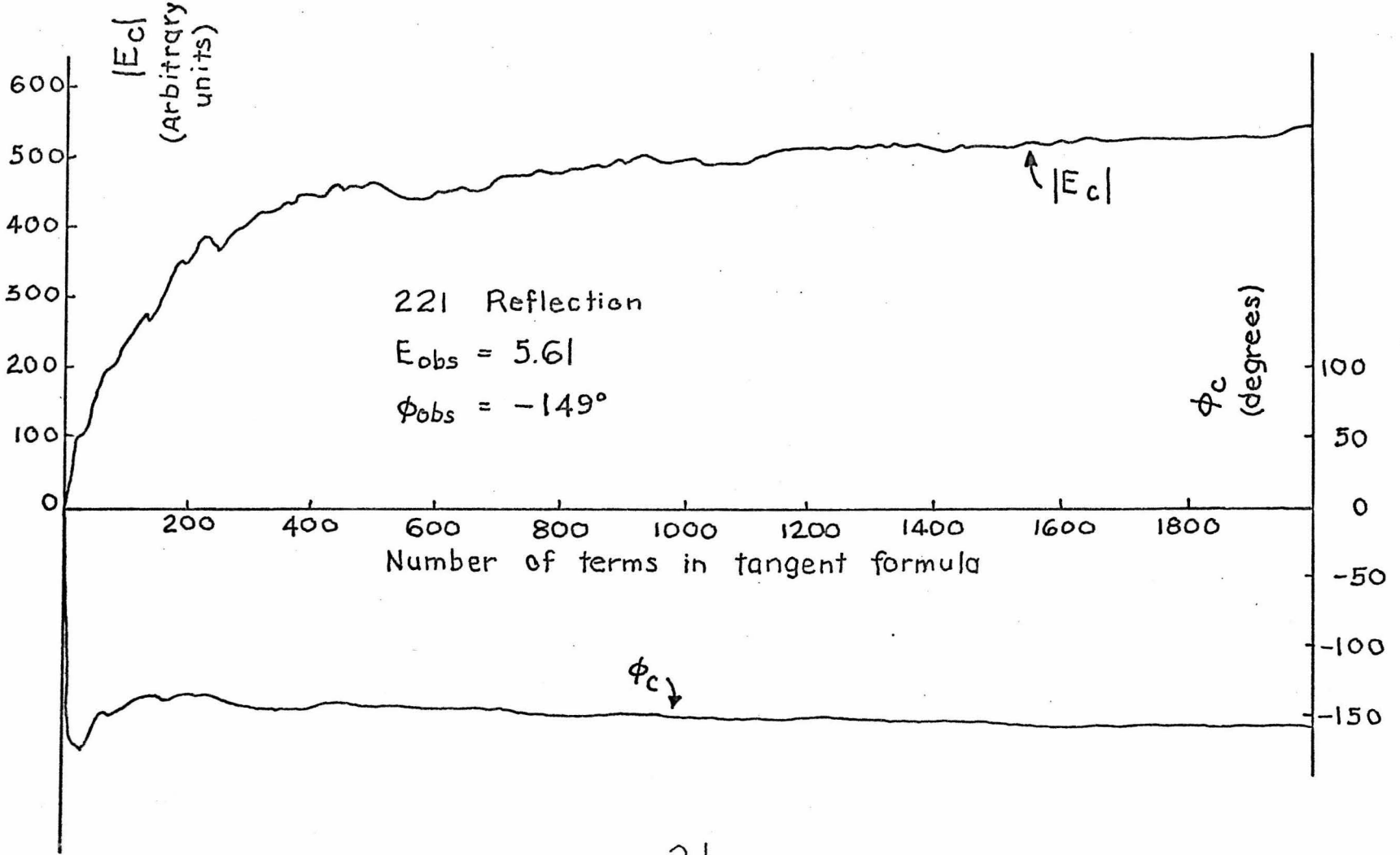
Figure 18. Map produced by using the joint probability distribution obtained from Pt SIR and tangent formula information.


70
60
50
40
30
20
10
0
1.0
2.0
3.0
4.0
5.0
5 A DATA
4 A DATA
⟨|φobs − φc|⟩ (degrees)
⟨EOBS⟩

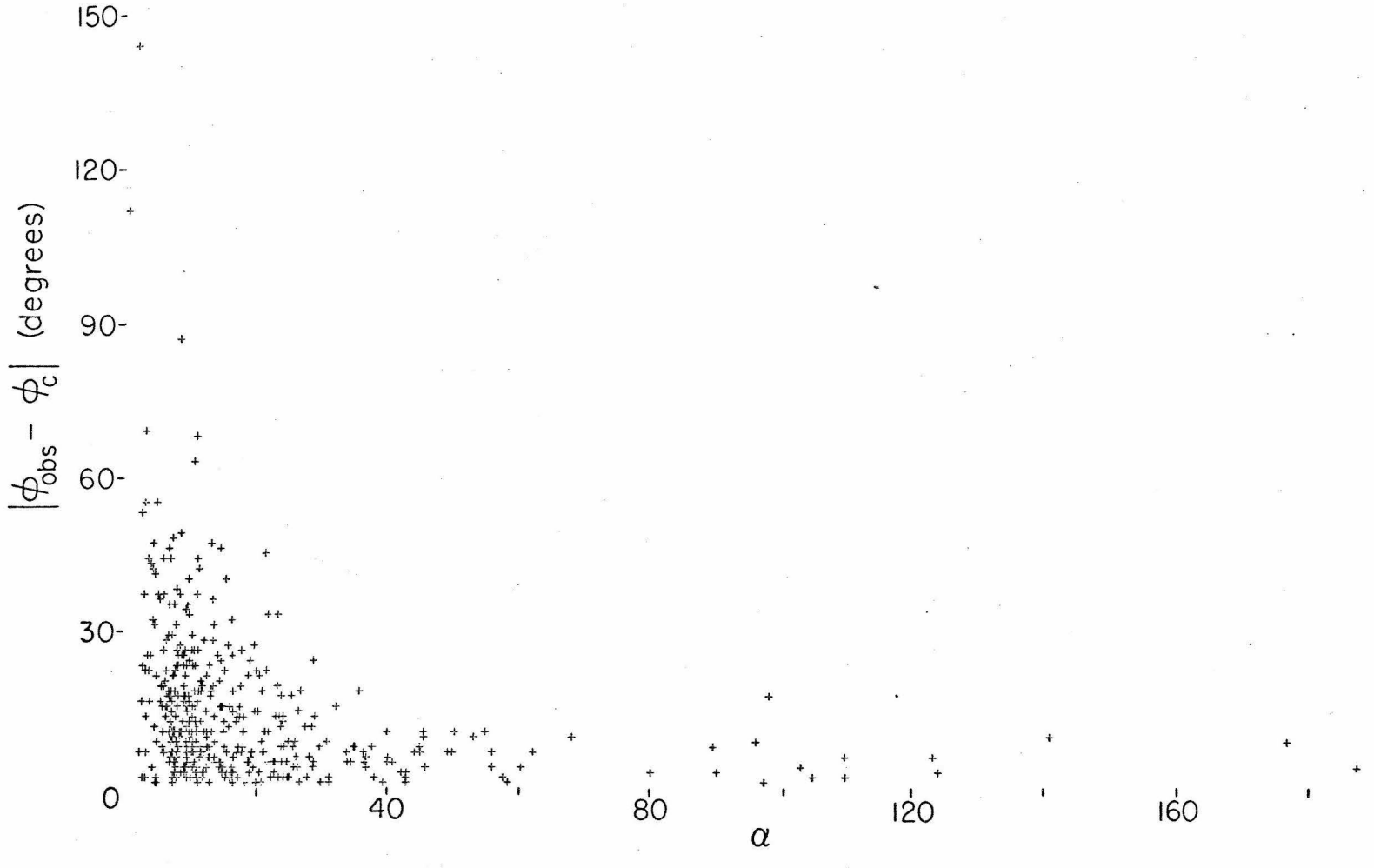
1.


|Ec|
(Arbitrary units)
500
400
300
200
100
0
002 Reflection
Eobs = 3.77
φobs = 0.0
|Ec|
φc
φc (degrees)
100
50
0
0
200
400
600
800
1000
1200
1400
1600
1800
2000
Number of terms in tangent formula

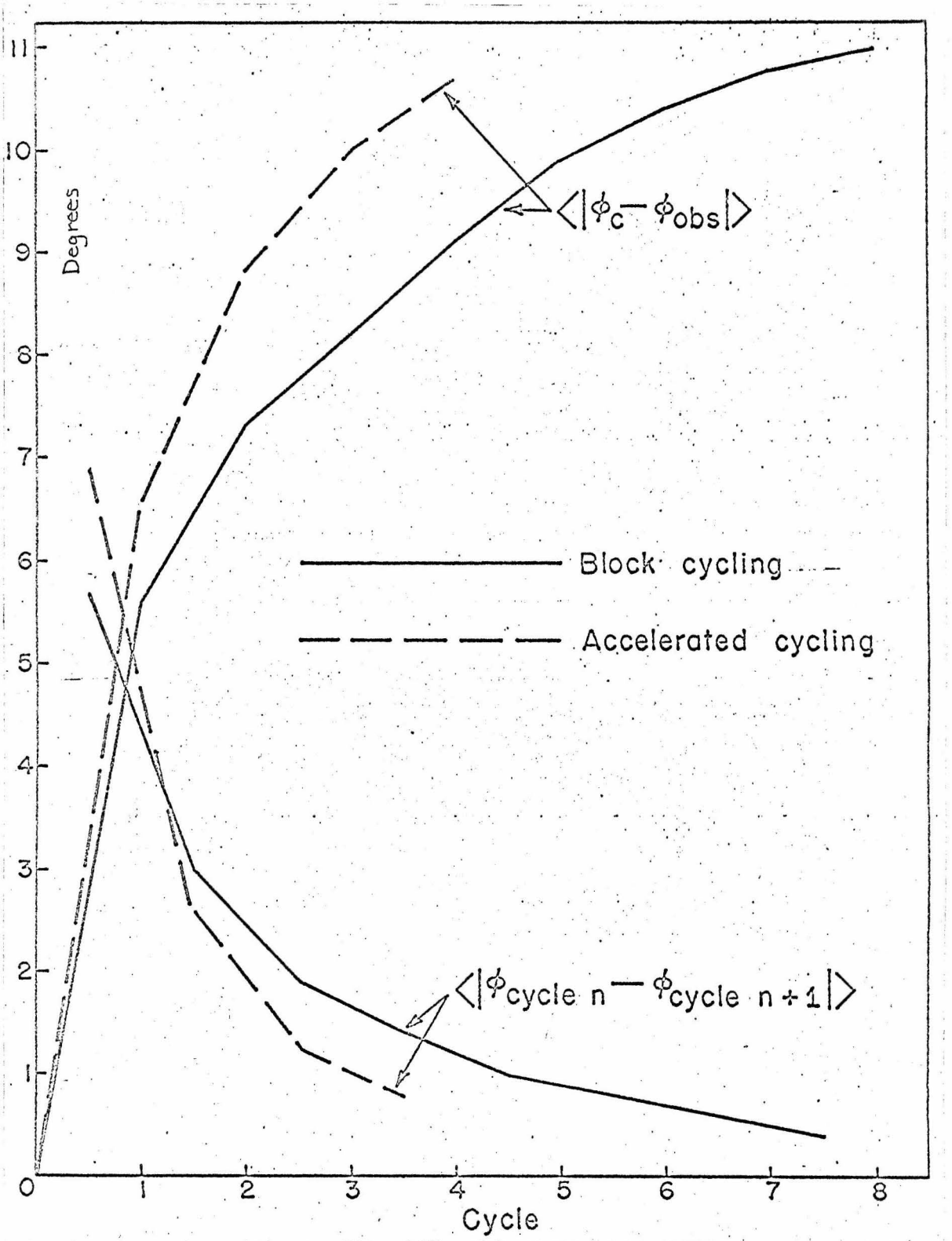
2 a.


|Ec|
(Arbitrary units)
600
500
400
300
200
100
0
|Ec|
221 Reflection
Eobs = 5.61
φobs = −149°
φc
(degrees)
100
50
0
-50
-100
-150
200
400
600
800
1000
1200
1400
1600
1800
Number of terms in tangent formula
φc

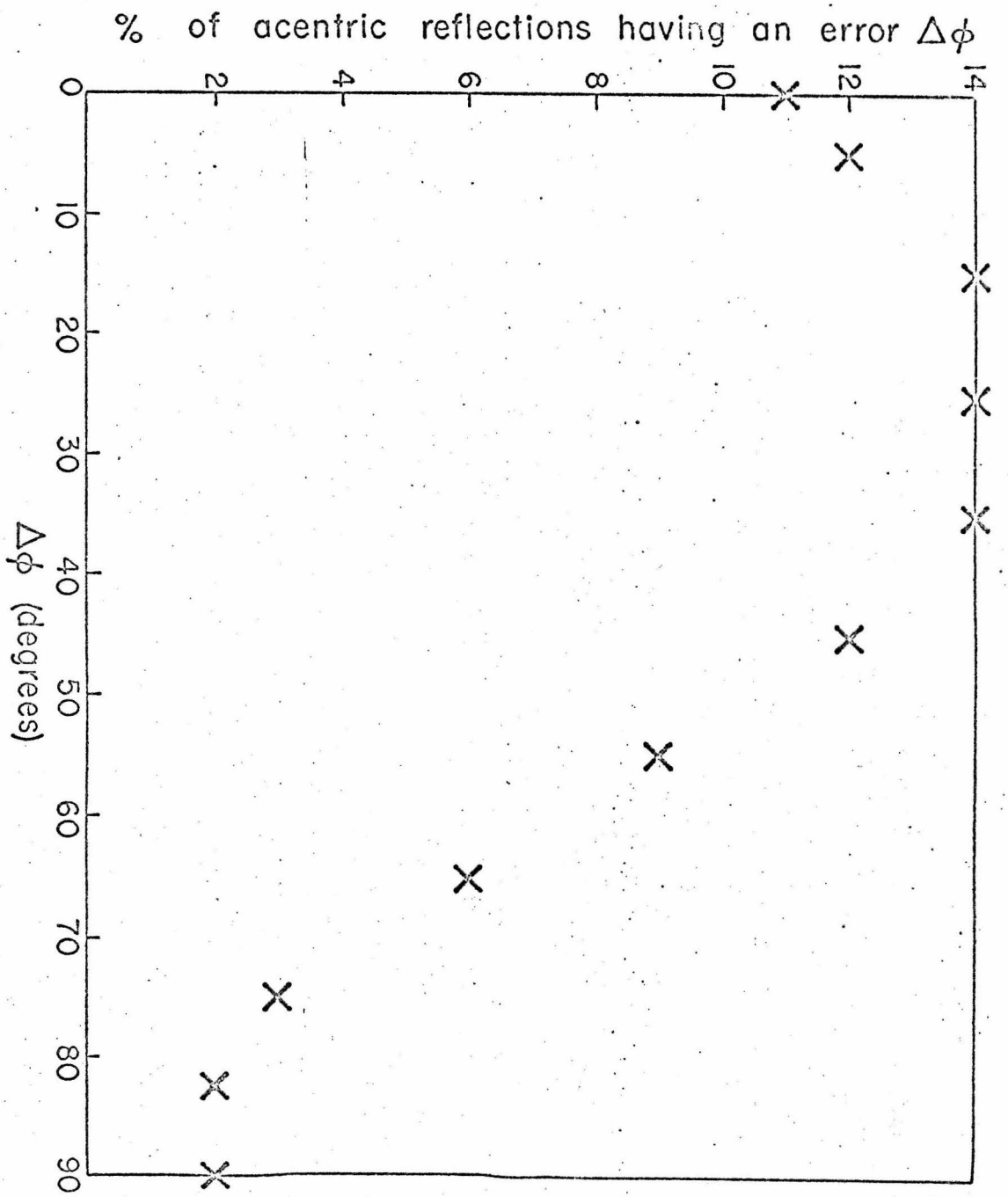
2b.



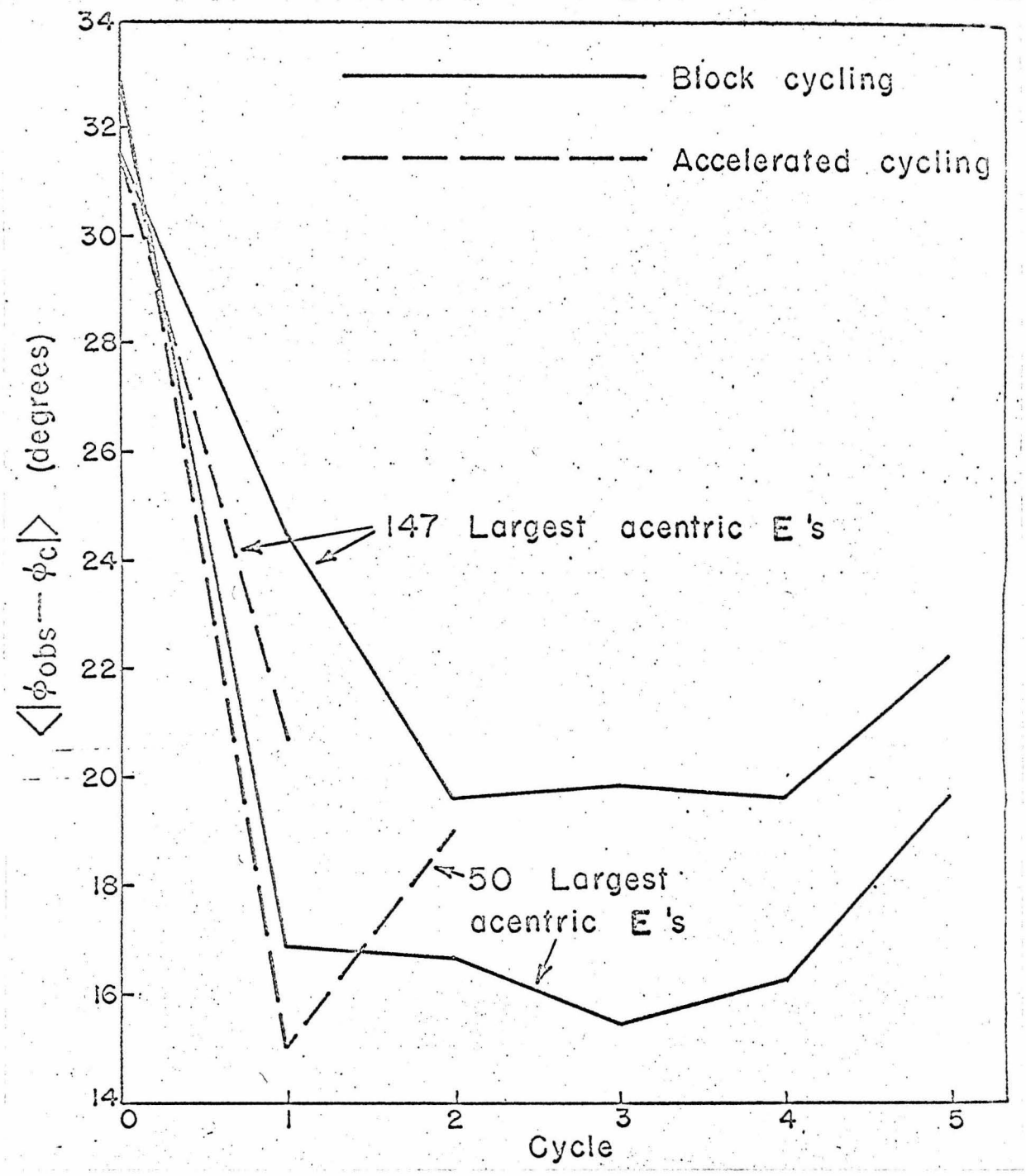

3.


Degrees
Cycle
$\langle|\phi_c - \phi_{obs}|\rangle$
$\langle|\phi_{cycle\ n} - \phi_{cycle\ n+1}|\rangle$
Block cycling
Accelerated cycling

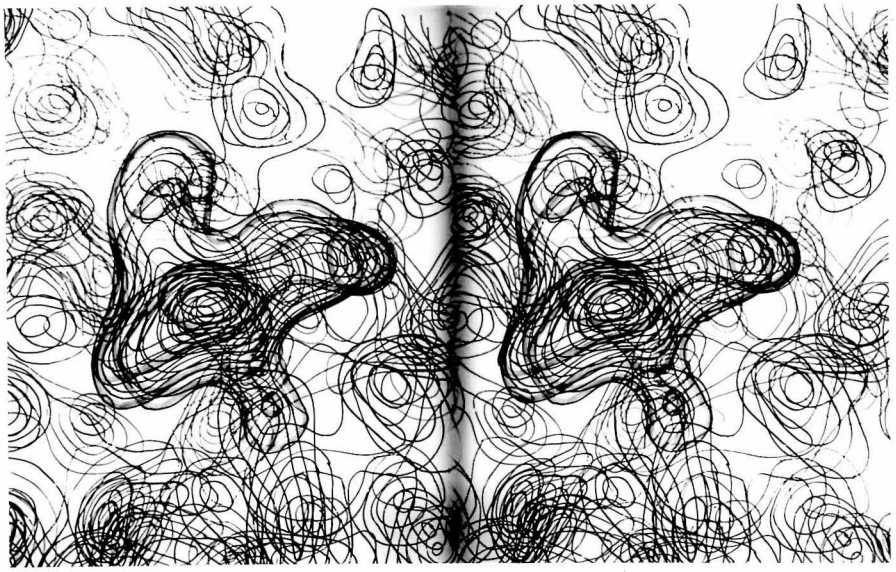
4.


% of acentric reflections having an error Δφ
0
2
4
6
8
10
12
14
Δφ (degrees)
10
20
30
40
50
60
70
80
90

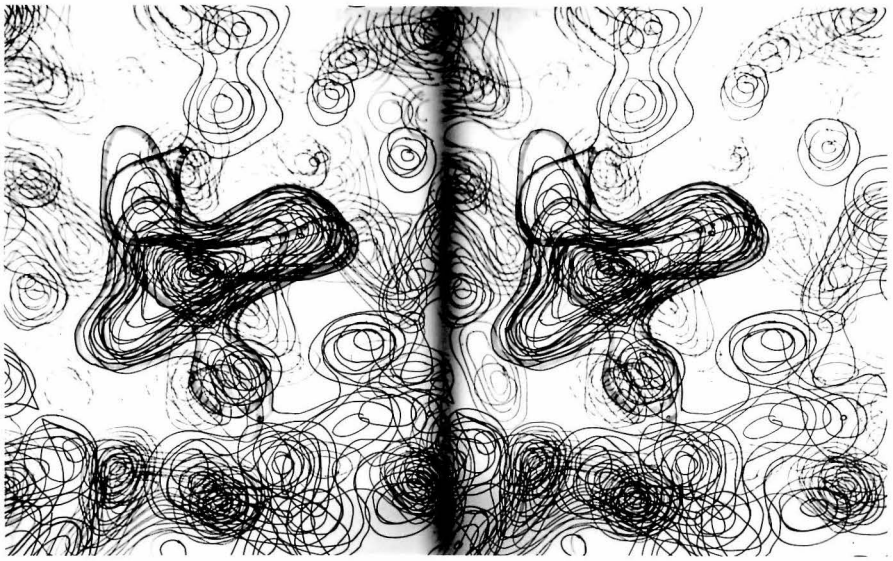
5.


Block cycling
Accelerated cycling
147 Largest acentric E's
50 Largest acentric E's
<|φobs − φc|> (degrees)
Cycle
34
32
30
28
26
24
22
20
18
16
14
0
1
2
3
4
5

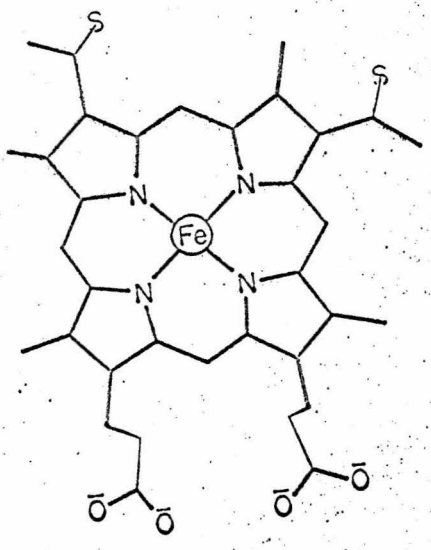
6.



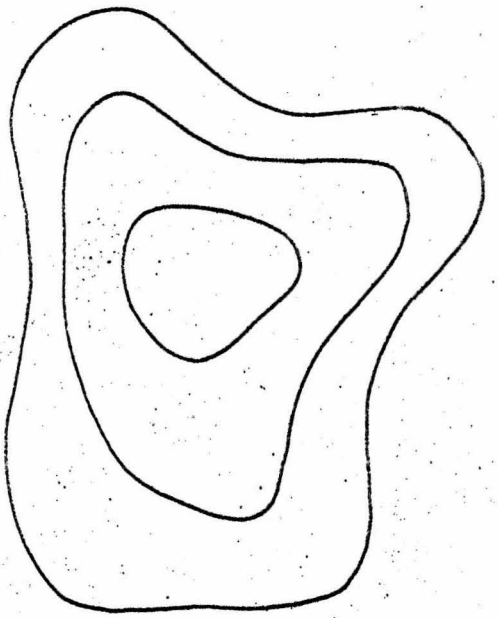
7.



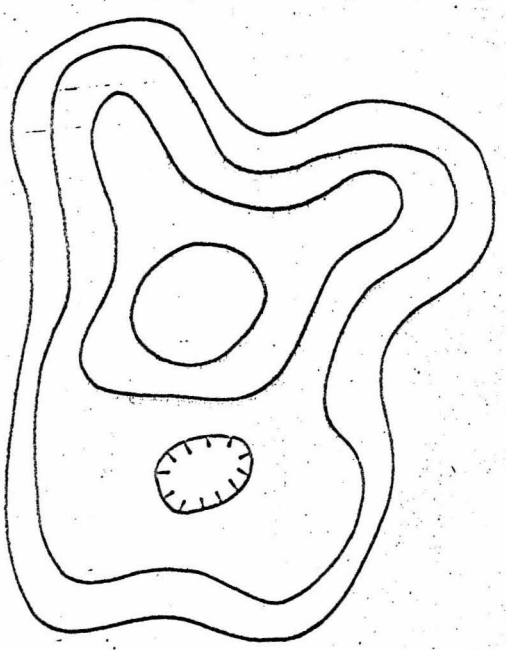
8.



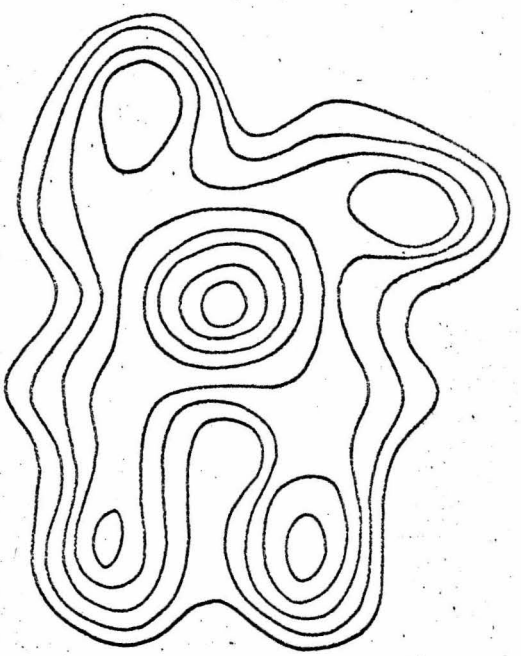

(a)



(b)

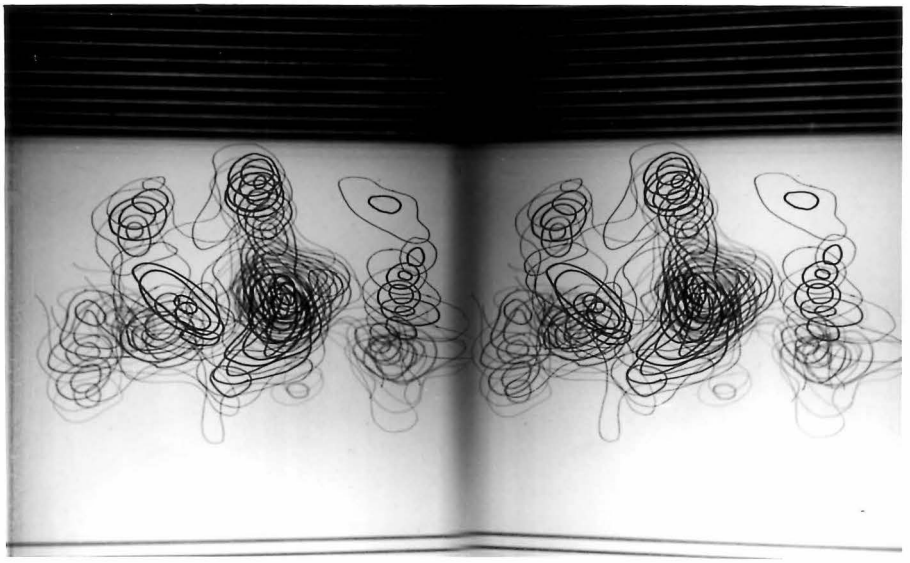


(c)

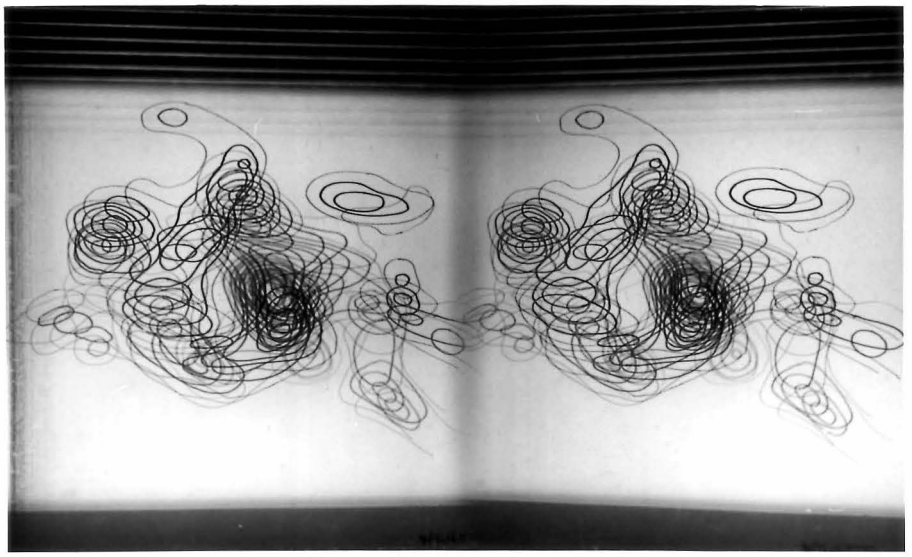


(d)

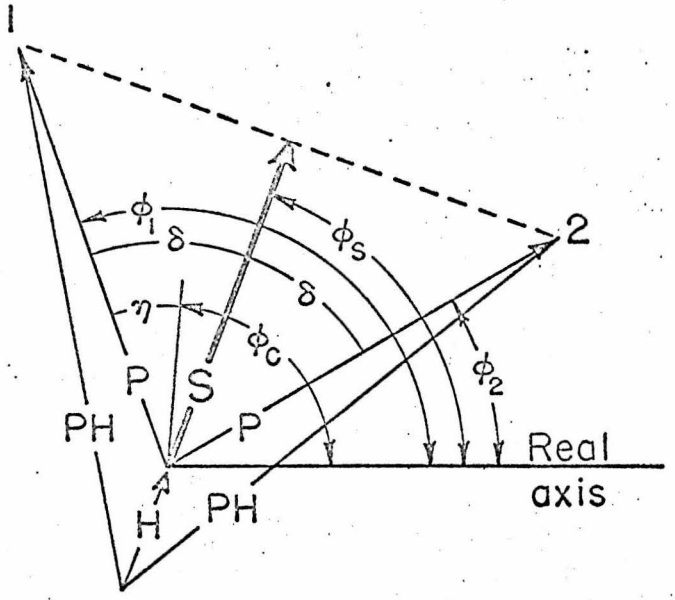
9.



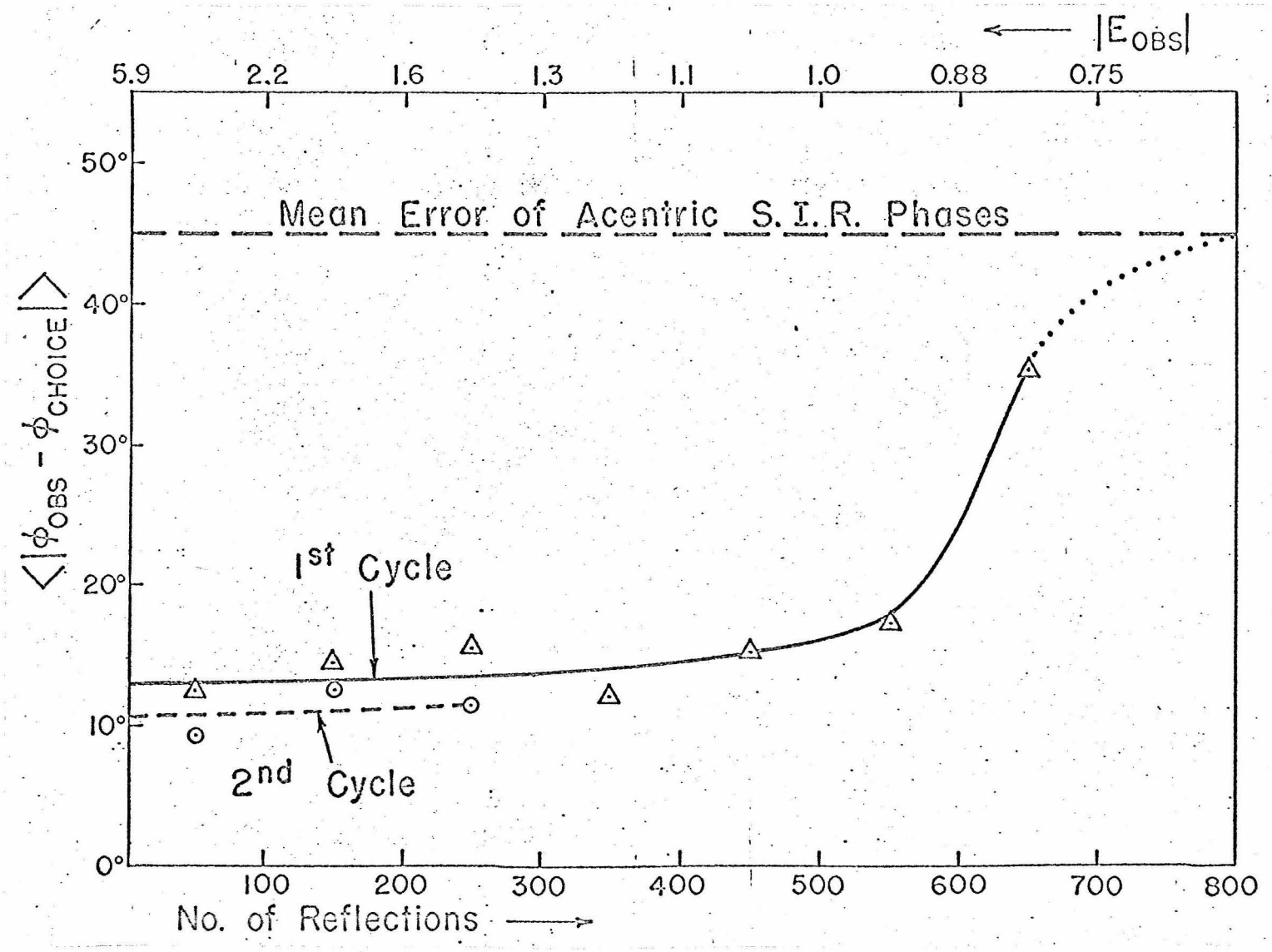
10.

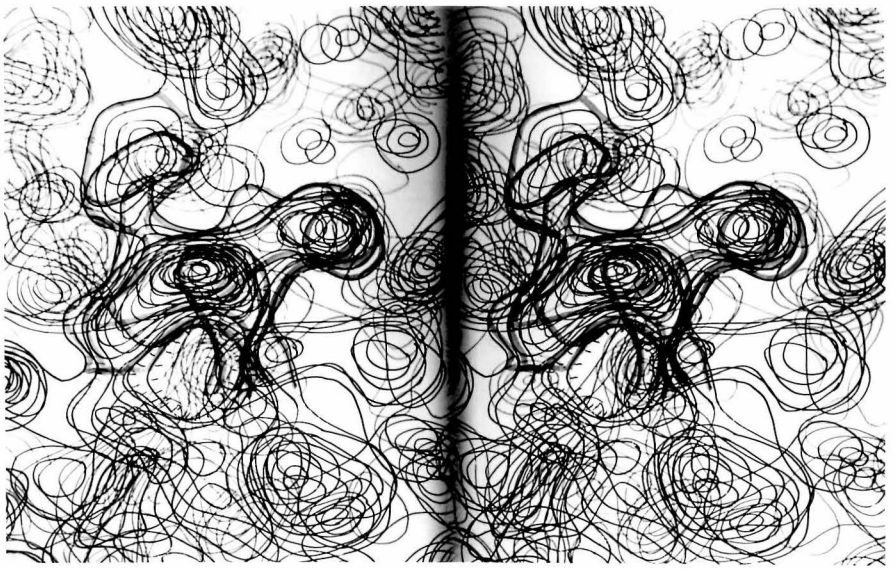


11.

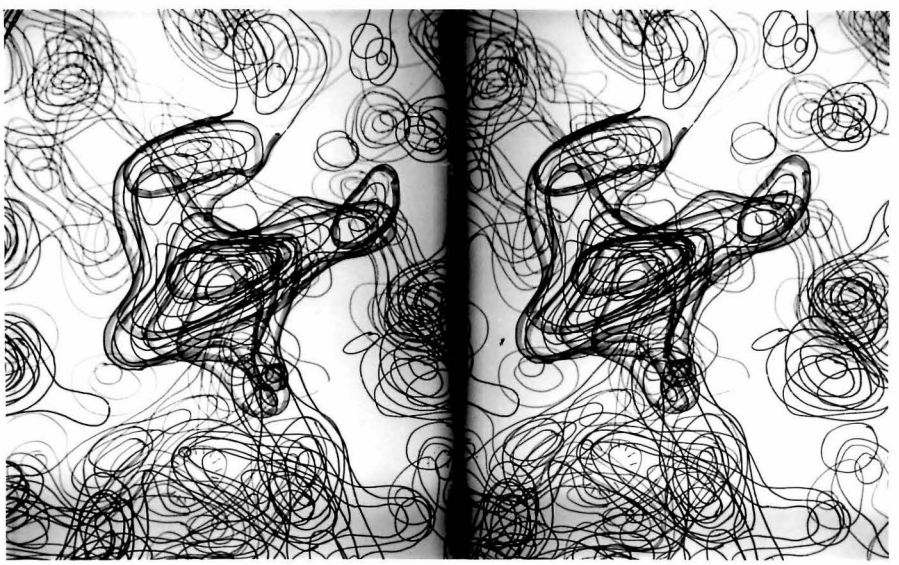



12.

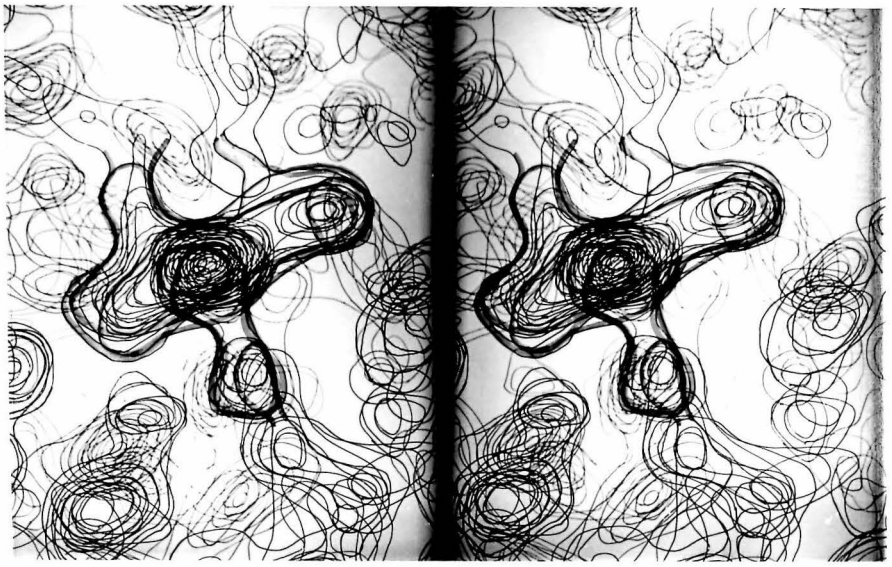

← |E_OBS|
5.9
2.2
1.6
1.3
1.1
1.0
0.88
0.75
50°
40°
30°
20°
10°
0°
Mean Error of Acentric S.I.R. Phases
⟨|φ_OBS − φ_CHOICE|⟩
1st Cycle
2nd Cycle
100
200
300
400
500
600
700
800
No. of Reflections →



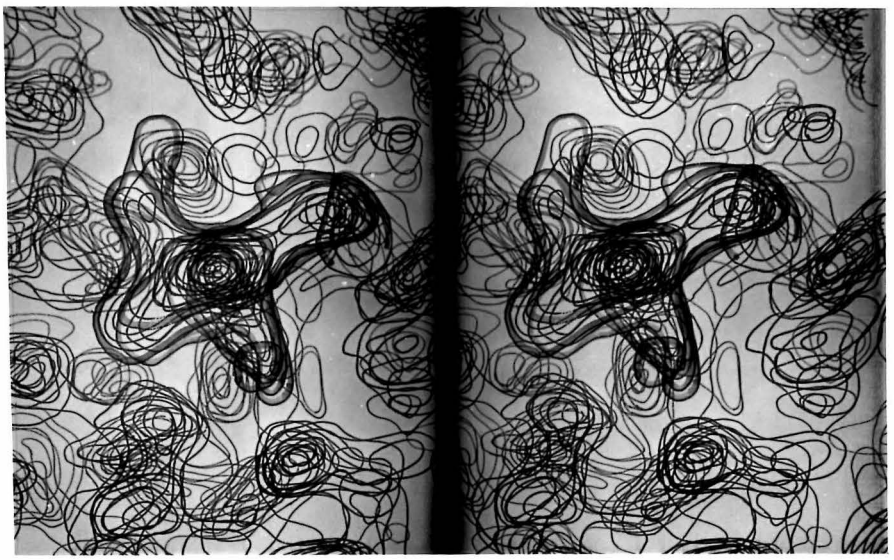
14



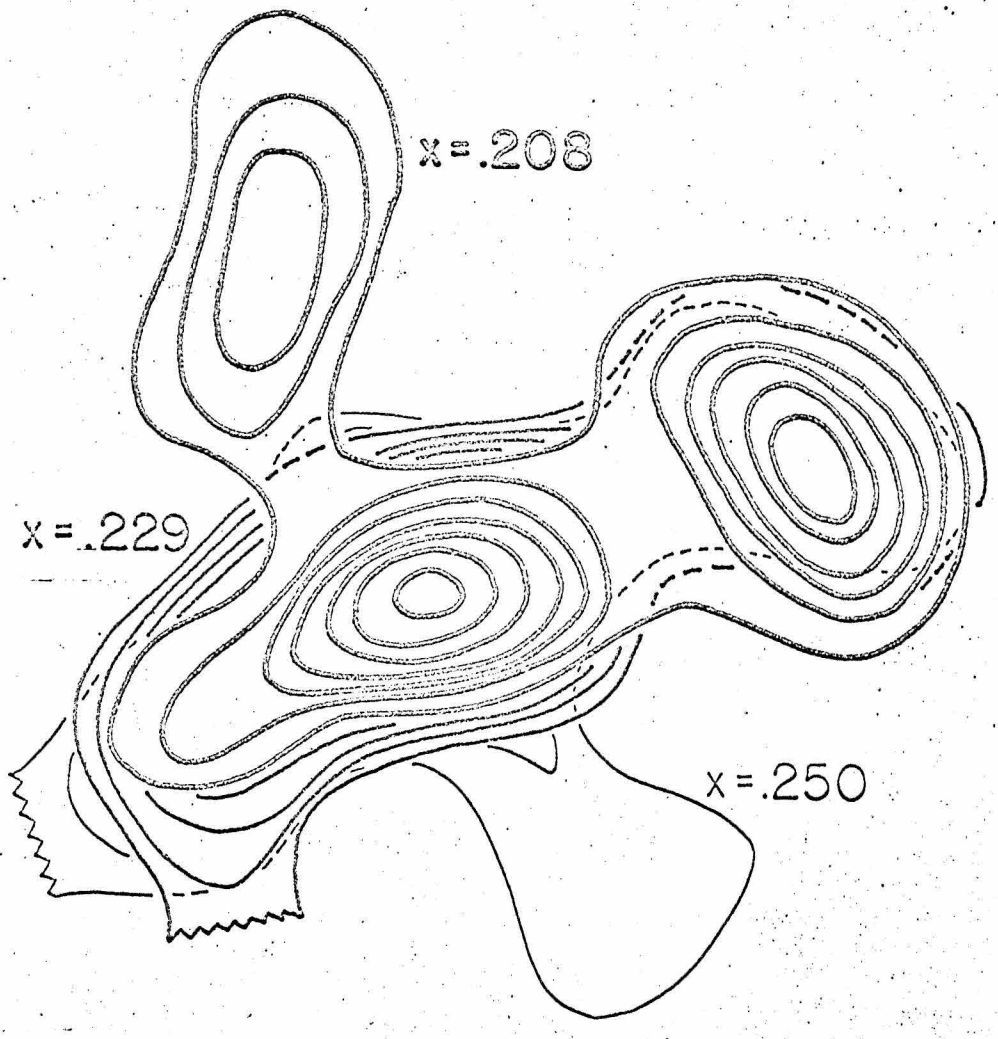
15



16.



17.




18.

## ABSTRACT OF PROPOSITIONS

1. In many laboratories, organic compounds, basic to life, have been abiogenically synthesized from mixtures of simple reducing gases believed to constitute the primitive atmosphere of the earth. Among these compounds are found most of the amino acids occurring in living systems today. The polymerization of these amino acid subunits into large protein molecules has not been observed in these experiments, however, probably because of the large energies necessary to initially synthesize the amino acids. A mechanism for thermal polymerization of amino acids based on the work of S. W. Fox is proposed.
2. Porphyrins and porphyrin-like compounds are distributed universally throughout the plant and animal kingdoms. Complexed with various proteins, porphyrins carry out two of the most fundamental life processes, oxidative phosphorylation and photosynthesis. An abiogenic origin of porphyrin-like compounds is proposed and its evolutionary implications are discussed.
3. Cytochrome c, being a rather loosely bound basic protein, would be ideally suited to act as a genetic regulator in the mitochondrion. A mechanism for regulating the binding of cytochrome c to the mitochondrial matrix is proposed and several possible effects of the "released" cytochrome are discussed.
4. The application of the Karle-Hauptman tangent formula to protein phasing has been discussed in Part II of the preceeding Thesis.

The specific application of the tangent formula to single isomorphous replacement phases is discussed, and a method for its application to a real data set is proposed.

5. The refinement of the atomic parameters for a protein structure is difficult due to a combination of the errors in the intensity data and the low ratio of observable intensities to unknown parameters. A least squares refinement method utilizing both intensity and phase information is proposed.

## PROPOSITION I

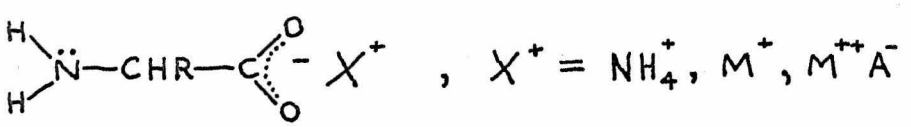
## A MECHANISM FOR THE THERMAL POLYMERIZATION OF AMINO ACIDS

The polymerization of amino acids into high molecular weight proteins under conditions presumed to exist on the primitive earth was first accomplished by S. W. Fox in 1954 (1). Since then, little has been said about this polymerization in terms of the actual reacting species of amino acids likely to occur during "early protein" synthesis.

The polymerizations were carried out by heating a mixture of amino acids anhydrously to a temperature of about 200 degrees. In order to get high molecular weight proteins, however, it was found that a large excess of aspartic or glutamic acid or lysine was necessary in the reaction mixture. When one of these three amino acids was not present in excess, analysis of the reaction mixture showed that a side reaction producing diketopiperazines predominated. The composition of the polymer formed when an excess of one of the three acids was present showed that the fraction of any given amino acid incorporated into the protein was directly proportional to its initial concentration in the reaction mixture. It was also implied by endgroup analysis that the sequence of amino acids was not entirely random. The molecular weights of the proteins produced by this method generally ranged from 4,000 to 9,000.

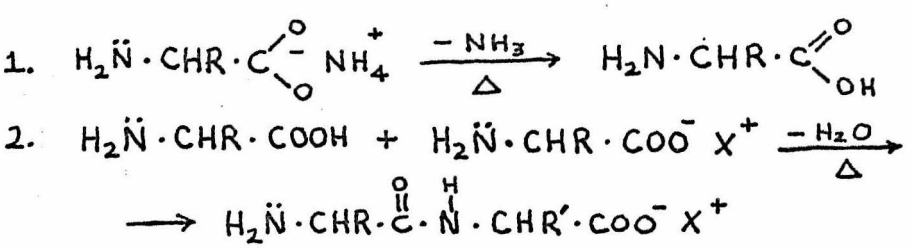
The most frequent criticism of Fox's polymerization is that his reaction conditions do not really represent those present upon the primitive earth. Many other compounds, as well as amino acids, are presumed to have been produced abiogenically and thus the polymerizations carried out on the primitive earth would have been done in the presence of large amounts of "contaminants" which could affect the polymerization. One would also expect that the amino acids would, at one time, have been dissolved in the sea and thus would be contaminated by a wide variety of cations and anions. In a recent paper (2), Fox has stated that if small amounts of inorganic phosphate are added to the reaction mixture, the reaction can be carried out at reduced temperatures with as good an efficiency as at a higher temperature in the absence of phosphate.

There has been much speculation about the composition of the atomosphere of the primitive earth. Almost all authors, however, believe that there was a large amount of ammonia present, at least in the probable period of abiogenic organic synthesis. This would insure a basic pH in the ocean, and a large concentration of ammonium ions. Thus, if an amino acid were transported to a region of thermal activity from the sea, it would be in the form:




Assuming regions of volcanic activity as the sites of thermal synthesis, two possible mechanisms could account for the polymerization of amino acids into long proteins. The amino acids, in their fully

ionized state could neither polymerize nor form diketopiperazines. If the amino acids were subjected to the proper thermal energies, however, an occasional ammonium ion could be converted into ammonia generating an active species of the associated amino acid. This should be immediately attacked by one of the fully ionized amino acids, since their amine groups would be strongly activated by the ionized carbonyl groups:



In this way, then, long proteins could be built up in the absence of the side reaction producing diketopiperazines.

An appreciable fraction of the fully ionized amino acids would be associated to a metal ion instead of an ammonium ion. In this case, protonation of an occasional amino acid carboxy group could occur by reaction with hydrogen sulfide, generally found in volcanic regions. Certain partially ionized anions such as $H_2PO_4^-$, probably present in the ocean at that time, could also serve as hydrogen ion donors to activate the amino acids. The major requirement for any activator of this nature is that the equilibrium concentration of the protonated amino acids be small under thermal conditions.

Addition of an amino acid to a growing polypeptide chain could occur in four ways under thermal conditions:

1. Activation of the C-terminus of the polypeptide followed by addition of an unactivated monomer.
2. Activation of the C-terminus of a monomer with subsequent addition to the N-terminus of the polypeptide.
3. Attack of an unactivated monomer at an internal position on the polypeptide.
4. Attack of an activated monomer at an internal position on the polypeptide.

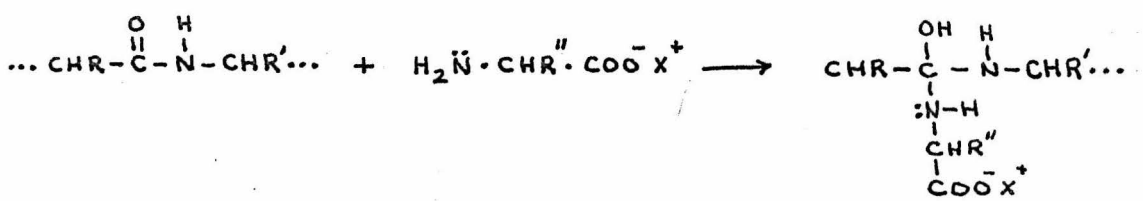
In mechanism 1., the rate of activation of the C-terminus of the polypeptide should be strongly affected by the nature of the R-group at the C-terminus. One might also expect a weak effect from the next internal amino acid. The unactivated monomers should all have about the same reactivity in this reaction since the presence of a full negative charge on the carbonyl group would mask any effect of the R-group. In terms of preferred sequences, amino acids whose R-groups tend to pull electrons away from the carbonyl group might tend to terminate the carboxy end of the polypeptide.

Addition to a growing polypeptide chain by mechanism 2. should depend upon the R-groups of both reacting species. The rate of activation of monomer will be a function of the monomer R-group, and, since in a polypeptide of several amino acids, the carbonyl group is effectively isolated from the N-terminus, the effects of R-groups close to the N-terminus of the polypeptide become important. Here, then, is a mechanism which could generate non-random sequences. The amino acid whose R-group placed the most positive charge on alpha-carboxy group

would be the most difficult to activate, since the activation involves protonating the carboxyl group, but this amino acid, once activated, would be more reactive than other amino acids to a nucleophillic attack since its carboxy group would carry more positive charge. If the difference between the reaction rates of the various protonated amino acids were great enough, however, there would still be a preference for the amino acid possessing an electron withdrawing R-group.

There are two possibilities for substitution of an amino acid at an internal position along a polypeptide chain. Since the amide linkages in a protein have a small amount of resonance energy, this type of substitution would not be expected to be important except at higher temperatures.

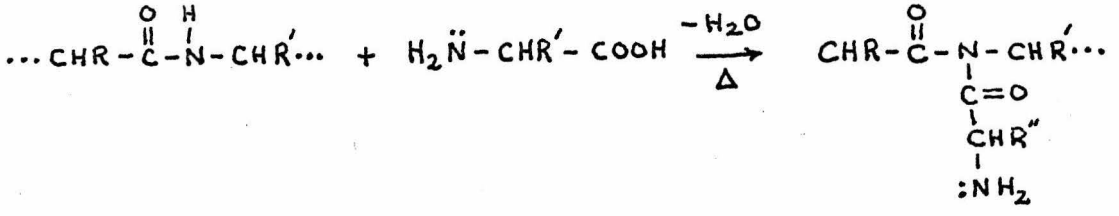
In mechanism 3., the first step involves the addition of a completely ionized monomer to an internal carboxy group of the polypeptide. This step would be reversible, and since part of the resonance energy of the polypeptide would be lost upon addition, the addition product would be present in very small quantities.



Before the added amino acid could be incorporated into the chain, its free carbonyl group would have to be protonated by one of the mechanisms mentioned previously. Since this mechanism depends upon two very unfavorable equilibriums, internal substitutions of this type

would probably be negligable.

A much more probable mechanism for internal substitution is mechanism 4. Here, the monomer must be activated before addition, but the addition would be irreversible:



The site of insertion along the polypeptide chain would be highly dependent upon R-group effects. Substitution would be favored by an electron donating R-group next to a backbone nitrogen. The effects of monomer R-groups would be similar to those discussed for mechanism 2.

Mechanisms such as 3. and 4. would be expected to predominate only after large polypeptides had been built up and the monomer concentration was fairly low. At this time, the monomers left would be of predominantly one type, that is, those whose reactivities were low for end addition. One might expect, therefore, that this type of amino acid would tend to follow a basic amino acid in the completed protein. Furthermore, this type of amino acid would tend to terminate both ends of the polypeptide. This result should be true for any mechanism of polymerization as long as there is any dependence of polymerization rate upon R-groups and might explain the apparent non-randomness of endgroups found by Fox in his thermal proteinoids.

## BIBLIOGRAPHY

1. Fox, S. W., Johnson, J. E., and Middlebrook, M., J. Am. Chem. Soc., 77, 1048, 1955.

2. Fox, S. W., Nature, 205(4969), 328-340, 1965.

Related References:

Fox: Arch. Biochem. Biophys., 109(1), 49-56, 1965.

Int. Congr. Biochem., 6(2), 155, 1964.

Int. Congr. Biochem., 6(2), 162, 1964.

Nature, 203(4952), 1362-1364, 1964.

Keosian, J., The Origin of Life, Reinhold Publishing Co., 1964.

# PROPOSITION II

# AN ABIOGENIC SYNTHESIS OF PORPHYRINS

Porphyrins and porphyrin-like compounds are distributed universally throughout the plant and animal kingdoms. Complexed with various proteins, prophyrins carry out two of the most fundamental life processes, oxidative phosphorylation and photosynthesis. An abiogenic origin of porphyrin-like compounds is proposed and its evolutionary implications are discussed.

Most authors agree that the primitive earth's atomsphere was composed of carbon, nitrogen, and oxygen in their reduced states, that is, methane ammonia, and water vapor (1). It has been experimentally verified that complex organic molecules can arise from such a mixture upon supplying the necessary energy of activation. Effective energy sources include short wavelength U.V. light, electrical discharge, high energy electrons, and thermal energy (under volcanic type conditions)--all agencies presumed to be active on the primitive earth (1). Since the porphyrins are so important to life processes, they must also have been formed by similar processes and not by a biochemical mechanism.

Several authors believe that the evolutionary reactions leading to a particular biological compound are preserved in the biosynthetic pathways of living organisms today. The current biosynthetic pathway

for the porphyrins utilizes succinic acid and glycine as precursors and through a series of condensations and oxidations produces the several porphyrins known in nature.

The first reaction of this sequence (Fig. 1) involves a condensation which would be very improbable without the assistance of a very specific enzyme. Since porphyrins are presumed to have evolved before the appearance of specific enzyme systems, the biosynthetic pathway does not appear to have preserved the abiogenic origin of the porphyrins.

A more likely abiogenic synthesis of porphyrins involves a condensation of alanine and a long-chain acid. (Fig. 2.) Since this sequence involves a loss of water in almost every step, one would expect the pathway to be favored by a dry, hot environment. Many authors have favored regions of volcanic activity as the sites for "early protein" synthesis (2). Since the environment around a volcano is generally of a highly reducing nature, further reactions between the substituted pyrrole rings would be unlikely. If these compounds were transferred to an aqueous medium, however, the methyl group could be oxidized to an aldehyde group (U.V. light supplying the necessary energy of activation). Two molecules of the monomer could then condense forming a dimer which would be highly stabilized due to internal hydrogen bonding.

The hydrogen atom involved in the hydrogen bond would immediately be replaced by a metal ion such as Fe(II). Since the concentration of metal ions would be much greater than that of dimer, no further

FIGURE 1

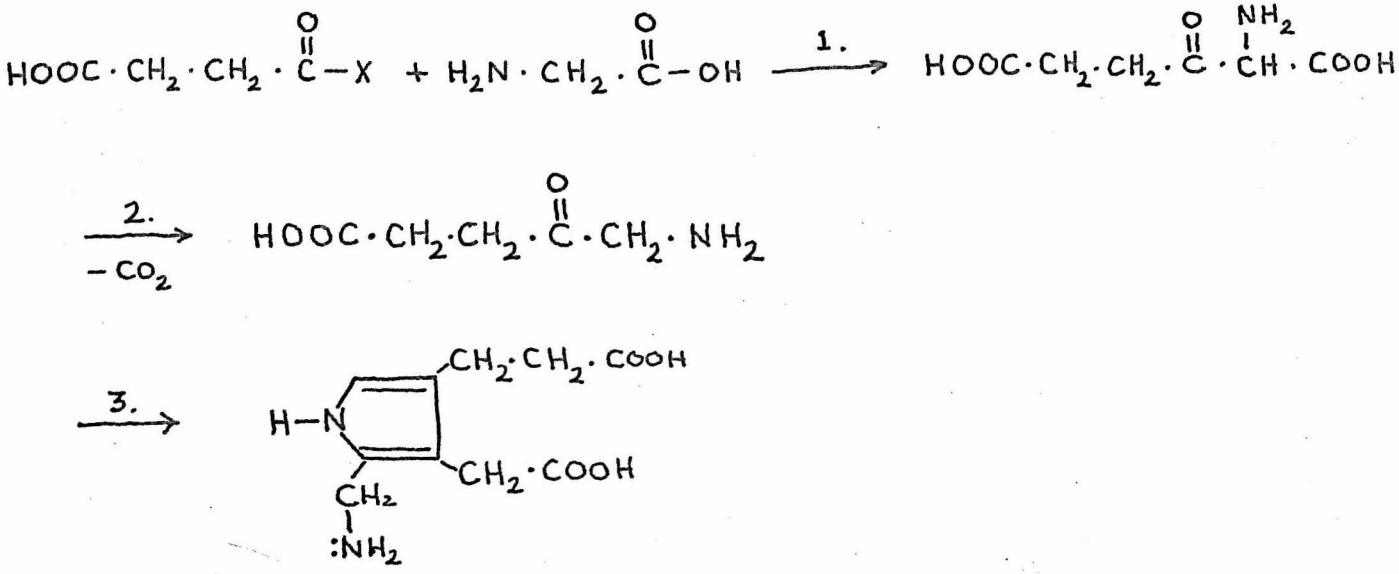
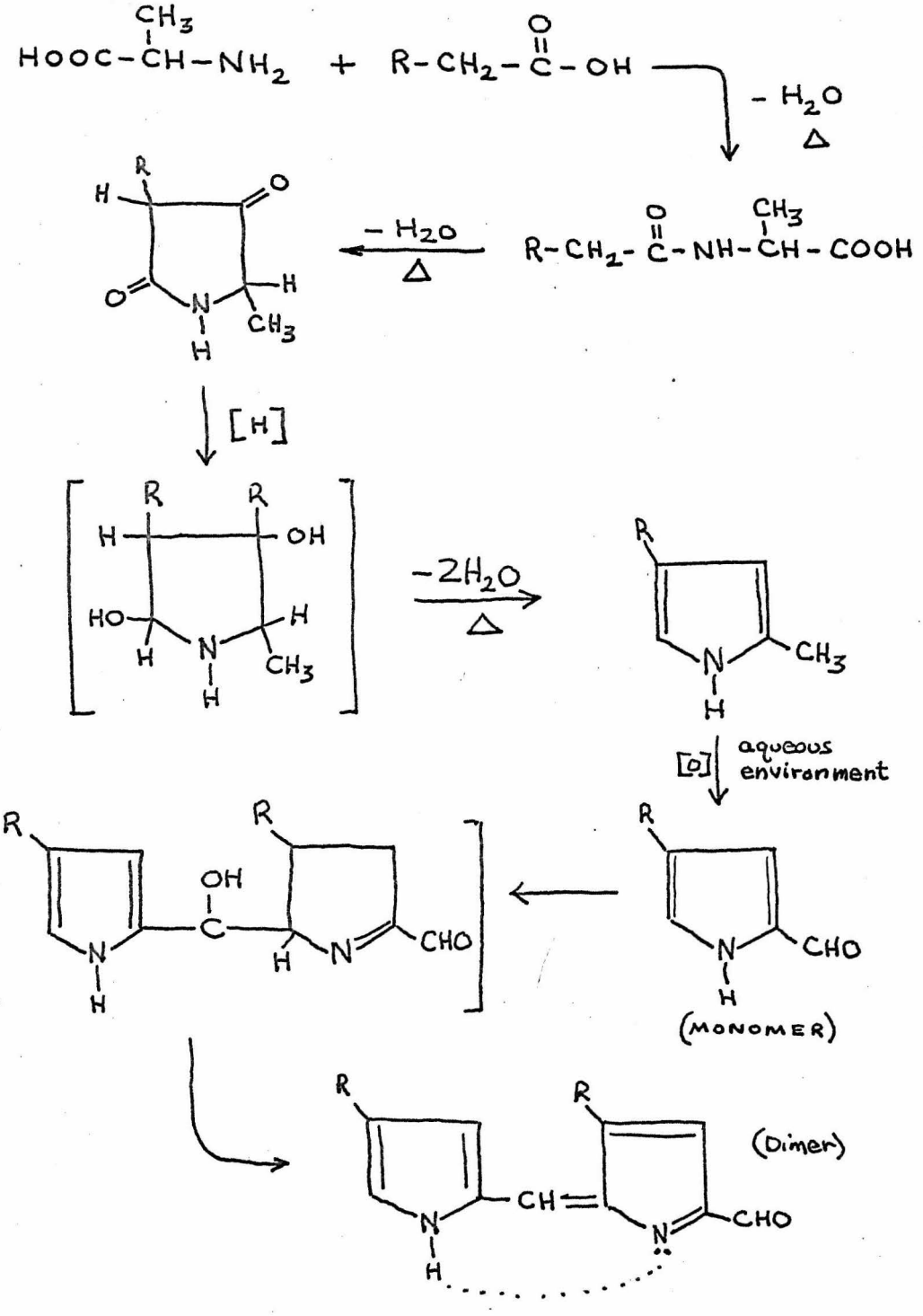

HOOC·CH2·CH2·C(=O)–X + H2N·CH2·C(=O)–OH
1.
HOOC·CH2·CH2·C(=O)·CH(NH2)·COOH
2.
–CO2
HOOC·CH2·CH2·C(=O)·CH2·NH2
3.
H–N
CH2·CH2·COOH
CH2·COOH
CH2
:NH2
4.
REDUCED PORPHYRINS
5.
[O]
PORPHYRINS

FIGURE 2


HOOC-CH-NH2
CH3
+ R-CH2-C-OH
O
- H2O
Δ
R-CH2-C-NH-CH-COOH
CH3
- H2O
Δ
[H]
-2H2O
Δ
[O]
aqueous environment
(MONOMER)
(Dimer)

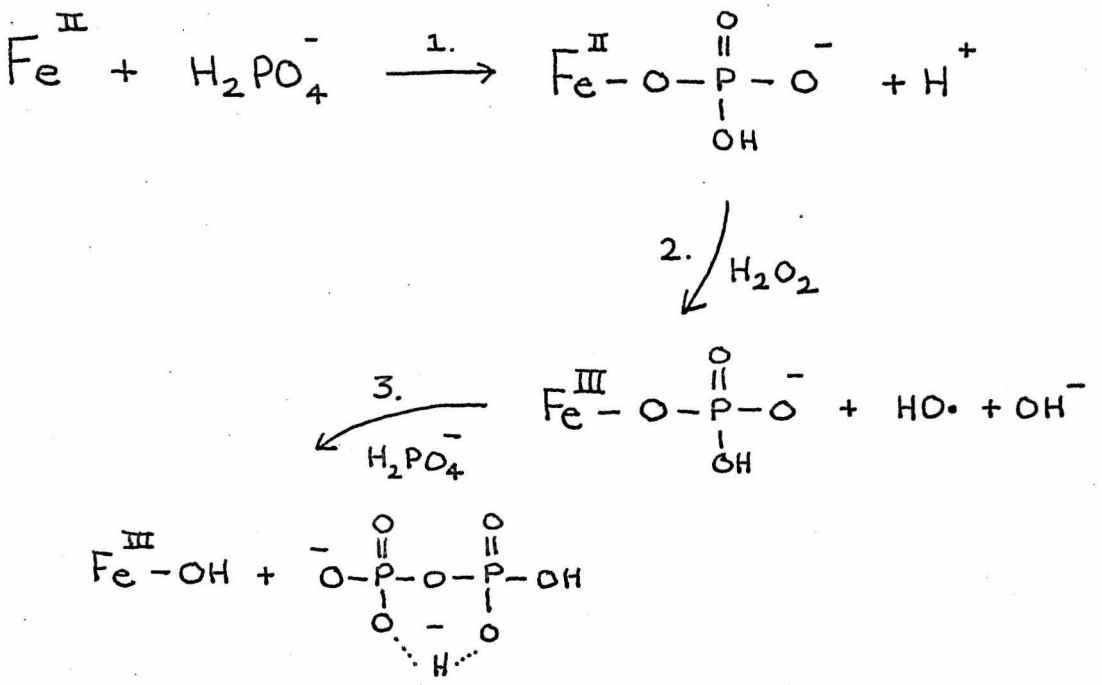
condensation of dimers should occur. For the production of a completed porphyrin compound some mechanism must evolve to concentrate the dimers.

It was mentioned before that sites of volcanic activity are believed to be the sites of early protein synthesis. If one dissolves several micrograms of proteinoid material prepared in this way, in the laboratory, in several ml. of hot water and cools the resultant solution, a large number of microspheres appear (3). These microspheres can collect more proteinoid-like material, grow in size, and break apart into two smaller microspheres, each capable of growth and division.

These microspheres should tend to concentrate any large non-polar molecule from an aqueous environment, including the dimer in question. They should also tend to reject small inorganic ions, so the porphyrin to metal ion concentration could be altered to favor polymerization to the porphyrin-like compound. Once formed, the porphyrin would be retained by the microsphere with numerous possibilities for histidines of the protein to occupy the fifth and sixth coordination positions of the metal ion.

It is known that if Fe(II) is oxidized by hydrogen peroxide in the presence of inorganic phosphate, a small amount of pyrophosphate is produced (4). (Fig. 3.) Surrounding the Fe(II) by a porphyrin ring should enhance this effect, and coordinating the porphyrin with a basic residue of a protein should enhance it even further. Thus, there was the possibility of a concurrent synthesis of protein and porphyrin-

FIGURE 3


Fe II + H2PO4 -
1.
Fe II - O - P - O - + H +
OH
2.
H2O2
Fe III - O - P - O - + HO• + OH -
OH
3.
H2PO4 -
Fe III - OH +
O - P - O - P - OH
H

precursor at the same site, insuring their presence together in the aqueous environment. The nature of the early protein insured a system which would concentrate organic molecules, grow, and divide. The presence of porphyrin precursor insured a method of breaking down hydrogen peroxide and storing the resultant energy in the form of high energy phosphate. The actual concentration of phosphate and synthesized pyrophosphate in the microspheres would be very small since they are small, highly charged inorganic ions. If a small concentration of organic phosphate existed in the droplet and participated in the reaction, however, much more of the energy produced could be retained by the microsphere.

It is impossible to determine exactly how any event in the past occurred. One can, however, show that a certain event was possible and even probable under a set of conditions presumed to exist at the time in question. Considering the proposed evolution of porphyrins, it must be shown that the chemical compounds needed for the monomer synthesis could have been available in large amounts, that the thermal condensations required to produce the monomer will occur under assumed volcanic conditions, and that the subsequent oxidation and formation of the dimer will occur under primitive-earth conditions. Assuming that the microdroplet theory is correct, it is also necessary to demonstrate that the microdroplets will concentrate the dimer and allow the final reaction leading to the porphyrin-like compound, and that the combination of "early protein" and porphyrin has a large catalytic effect upon the formation of pyrophosphate.

The formation of large amounts of alanine under primitive earth conditions has been demonstrated using several of the above mentioned energy sources. There has been no report of long chain acid synthesis under the conditions tested, however, many authors have postulated the presence of long chain hydrocarbons formed through free radical processes in the atmosphere. At that time, ultraviolet light continuously produced hydrogen peroxide and oxygen in the upper atmosphere and the presence of long chain acids is, therefore, possible.

The feasibility of the thermal condensation of alanine and a long chain acid can be directly tested in the laboratory. Salts of alanine and a variety of long chain acids can be placed in one of the proposed primitive-earth atmospheres and heated at moderate temperatures. The presence of the pyrrole nucleus in the reaction mixture can be determined by standard chemical tests, or by spectroscopy. The presence of porphyrin should also be tested for in this mixture since a polymerization of the monomer might occur under these conditions.

The formation of dimer and, or porphyrin in an aqueous medium from the monomer can be tested by similar experiments. The effect of the R-group on the reactivity of the proposed monomer should be small unless there is a site of oxidation close to the pyrrole nucleous. One can, therefore, use any available substituted compound as a model.

The effect of the microdroplet-porphyrin complex (if it exists) on the Fe(II)-HOOH-phosphate reaction is the most interesting aspect of the proposed evolution of porphyrins. To test the feasibility of this

catalysis, the efficiency of pyrophosphate formation compared to total Fe(II) oxidized must be measured for Fe(II) alone, porphyrin alone, and porphyrin plus microsphere. The pyrophosphate formed can be measured using standard chemical tests. The final Fe(II)/Fe(III) ratio can be determined spectrophotometrically. Assuming the reaction rate for porphyrin plus microsphere is much higher than Fe(II) alone, the porphyrin plus microsphere reaction can be run in the presence of an excess of Fe(II) to keep the porphyrin iron in its reduced state. Various organic phosphates can be added to the inorganic phosphate used initially to test their participation in the phosphate reaction.

Assuming the next step in the evolution of the microspheres would be a mechanism to polymerize amino acids found free in solution, small amounts of the acids can be added to the pyrophosphate-producing microspheres to test their incorporation into protein. Using radioactive tracers, the incorporation could be directly determined by means of chromotography.

## BIBLIOGRAPHY

1. Keosian, J., The Origin of Life.

2. Fox, S. W., Johnson, J. E., and Middlebrook, M., J. Am. Chem. Soc., 77, 1048, 1955.

3. Fox, S. W., and Yuyama, S., Ann. N. York Acad. Sci., 108, 487, 1963.

4. Calvin, M., in Horizons in Biochemistry, Academic Press, 50, 1962.

## PROPOSITION III

## THE BINDING OF CYTOCHROME C TO THE MITOCHONDRIAL MATRIX

Cytochrome c is the only protein in the cytochrome chain that can be easily isolated in a pure form. The other cytochromes of the chain seem to be tightly bound to the mitochondrial matrix, and their preparations are almost always contaminated by varying amounts of lipoprotein. In terms of the theory of selective adaptation, this "loose" binding of cytochrome c must either confer a selective advantage upon the organism involved or must pose no selective disadvantage to organisms which have evolved a tightly bound cytochrome c. In view of the importance of the cytochrome system in life, as it exists today, and the universality of "loose" cytochrome c, there must be some selective advantage in an organism possessing this property. A mechanism for the release and recapture of cytochrome c from the mitochondrial matrix is proposed and possible selective advantages are discussed.

An insight into the type of binding involved in the attachment of a cytochrome c molecule to the mitochondrial matrix can be obtained by looking at the isolation procedures used in the initial extraction of cytochrome c from tissue homogenates. The three most commonly used methods are extractions with solutions of: (1). neutralized trichloroacetic acid, (2). ammonium sulphate, (3). and aluminum

sulphate (1). All three methods of isolation employ solutions of high ionic strength. This indicates that the binding involved in the cytochrome c--mitochondrial matrix complex is essentially ionic in character. A logical mechanism for regulation of the binding constant of cytochrome c to the mitochondrial matrix, then, would be a variation in the ionic strength of the mitochondrial protoplasm.

Systematic variations of ionic strength of the mitochondrial protoplasm have been observed by many workers (2). It has been shown that mitochondria undergoing active oxidative phosphorylation actively transport many inorganic cations across the mitochondrial membrane from the surrounding cellular protoplasm. Upon cessation of oxidative phosphorylation, either by inhibition of electron transfer or by inhibition of phosphorylation by uncoupling agents, the ion gradient is lost. Thus, depending upon the environment of the mitochondria, the ionic strength of the mitochondrial protoplasm varies between the ambient ionic strength of the cellular protoplasm and a much higher value which depends upon the rate of oxidative phosphorylation. If cytochrome c is to be lost from the mitochondrial matrix under one of these extremes of ionic strength, it would most probably be lost under conditions of low ionic strength.

Lemly (3) has shown that if rat heart tissues are subjected to experimental anoxia, the cytochrome c content of these tissues decreases by about 30%. It has also been shown by Dutkiewick (4), that the oxygen consumption of guinea pig brain slices, when poisoned by a low carbon monoxide tension, could be partially restored by the

addition of a limited amount of cytochrome c. Both anoxia and carbon monoxide poisoning are conditions which limit the amount of oxidative phosphorylation in an otherwise normal cell. The loss of cytochrome c from the mitochondrial matrix in either case can be accounted for by by the change in the mitochondrial ionic gradient caused by the loss of oxidative phosphorylation.

Under conditions of extreme anoxia, it would be advantageous for a cell to be able to "turn off" most of its inessential metabolic processes. Cytochrome c, being a very basic protein, is ideally suited to act as a genetic regulator as is another type of basic protein, histone. Recent work by Olivera (5) at Caltech has shown that cytochrome c does indeed bind very well to DNA. The binding constant, however, is strongly dependent upon the ionic strength of the medium, as would be expected for a positively charged protein binding to a negatively charged DNA molecule. A drop in ionic strength of the mitochondrial protoplasm could, then, change the cytochrome c distribution within the mitochondrion between mitochondrial matrix protein and mitochondrial DNA (and possibly RNA).

The regulation of the mitochondrial DNA by cytochrome c could be either specific to a small number of genomes or could be general, producing a general shutdown of mitochondrial protein synthesis. Except in special mutants such as the poky mutant of Neurospora or the petite mutant in yeast, one would expect the regulation by cytochrome c to be specific since it is hard to imagine maintaining the production of essential metabolites and proteins under conditions

of a general metabolic shutdown. In the case of the mentioned mutants, the cytochrome systems are deficient and the organisms produce large amounts of extra-mitochondrial cytochrome c. Since the ionic strength of the cellular protoplasm is quite low (as compared to a mitochondria), one might expect non-specific binding of cytochrome c to ribonucleotides in the cellular protoplasm such as messenger RNA.

A second function of the DNA--cytochrome c complex could be to facilitate the removal of toxic substances such as carbon monoxide or cyanide from the mitochondrion. Margoliash (6) and others have shown that modified forms of cytochrome c, such as dimers and higher polymers, exhibit auto-oxidizability and carbon monoxide reactivity. If the binding of cytochrome c to DNA caused tertiary modifications in the cytochrome similar to those in the dimer, the presence of DNA--cytochrome complexes would tend to shift the cyanide or carbon monoxide molecules from the cytochrome oxidase to the complex. Since the binding of cytochrome to DNA is an equilibrium process, the effect of cytochrome-DNA complex would actually be to shift the carbon monoxide or cyanide from the cytochrome oxidase to the mitochondrial protoplasm (unmodified cytochrome c does not bind carbon monoxide or cyanide).

A third similar effect could arise if the cytochrome c were rapidly exchanged between the mitochondrial matrix and the mitochondrial DNA. Reduced cytochrome leaving the matrix could be oxidized at the DNA and then returned to the matrix, effectively bypassing the metabolic block at cytochrome oxidase.

The effect of mitochondrial ionic strength upon binding of cytochrome c to the mitochondrial matrix can be easily determined using radioactive tracers. Uniformly labeled cytochrome c can be obtained from a culture grown in the presence of labeled amino acids. The binding of this labeled cytochrome to isolated, intact mitochondria can then be measured at a number of different ionic strengths by spinning down the mitochondria after equilibrium has been reached and measuring the amount of radioactivity vs. total protein in the mitochondrial pellet. The equilibration must be done under anaerobic conditions or better in the absence of any reduced substrate to preclude any ionic gradient across the mitochondrial membrane due to oxidative phosphorylation.

After it has been shown that binding of cytochrome c to the mitochondrial matrix is a function of ionic strength, it is necessary to show that the ionic gradients produced by active oxidative phosphorylation are large enough to cause a change in the cytochrome binding constants. In this case, the labeled mitochondria must be prepared under conditions of oxidative phosphorylation since all measurements of the cytochrome equilibrium will be made upon respiring mitochondria in a solution whose ionic strength is similar to the cellular ionic strength. This can be achieved by first equilibrating the mitochondria with labeled cytochrome under anaerobic conditions in the low ionic strength buffer to be used in the final measurements. After equilibrium has been reached, the mitochondrial preparation can be made aerobic (to "fix" the bound cytochrome) and the mitochondria

separated from the excess, labeled cytochrome by centrifugation or by dialysis. The loss of labeled cytochrome by these mitochondria can now be measured by subjecting the mitochondria to brief periods of anaerobic conditions and measuring the bound cytochrome as before.

The binding of cytochrome c to mitochondrial DNA under anaerobic conditions will be much more difficult to demonstrate. Two methods can be used, one to try to observe the effect of exogeneous cytochrome on the synthesis of proteins known to be cytoplasmically inherited in whole mitochondrial preparations and the other to look for regions of high cytochrome concentration in mitochondria labeled with radioactive cytochrome and subjected to anaerobic conditions in a buffer of very low ionic strength.

Mitochondria are known to incorporate labeled amino acids into protein. The rate of incorporation, however, is low and there has been difficulty locating the mitochondrial proteins actually incorporating the label. The best method of showing any effect of cytochrome upon this synthesis would probably be just to measure the rate of incorporation of label, regardless of the proteins involved. In the case that mitochondrial DNA produces repressors or activators which act upon nuclear DNA, an effect of cytochrome c upon the repressor or activator site would be observable in whole cell cultures.

Assuming that at very low ionic strengths cytochrome c binds very weakly to the mitochondrial matrix and very strongly to the mitochondrial DNA, the location of the DNA in the mitochondrial structure could be determined using autoradiography and thin sectioning. The ionic

strengths used in constructing the labeled DNA must be much lower than the ambivalent ionic strength of the cellular protoplasm, since the presence of labeled cytochrome on the mitochondrial matrix would obscure the cytochrome--DNA complex.

## BIBLIOGRAPHY

1. Dr. E. Margoliash, unpublished results.

2. Lehninger, Albert L., The Mitochondrion, Benjamin, 1964.

3. Lemley, J. M. and Meneely, G. R., "Effects of Anoxia on Metabolism of Myocardial Tissue, Am. Jour. Physiol, 169(1), 66-73, 1952.

4. Dutkiewick, J. S., "The Effect of Cyt. c. on the Oxygen Consumption of Tissues of Normal and Carbon Monoxide poisoned Animals", J. Physiol., 152(3), 482-486, 1960.

5. Dr. Baldomero Olivera, personal communication (unpublished).

6. Margoliash, E., "Some Structural Aspects of Cytochrome c Function", Brookhaven Symposia in Biology: No. 15 (1962).

## PROPOSITION IV

## THE APPLICATION OF THE KARLE-HAUPTMAN TANGENT FORMULA TO SINGLE ISOMORPHOUS REPLACEMENT PHASING

As was stated In Part II of the preceeding Thesis, the most successful application of the Karle-Hauptman tangent formula to protein phasing has been its use as a method for selecting the correct phase from the two possible phase choices resulting from a single isomorphous replacement (SIR) phase analysis. The only theoretical limitation to this method is that the atomic cluster of heavy atoms in the unit cell be non-centric. This application of the tangent formula is an important advance in protein crystallography. With the tangent formula to break the SIR phase ambiguity, a protein structure could, in theory, be solved with a single heavy atom derivative. It is important, therefore, to determine the actual power of the method and the rules for its application to protein data sets in general.

The process of selecting one of the two possible SIR phases by this method consists of two completely independent steps. The first step is simply the calculation of a new set of phases using the tangent formula. The input phases to the tangent formula in this step can be either the SIR centroid phases or the SIR phases themselves, choosing one of the two possible phases for each reflection in a random manner. In the second step, the phase calculated by the tangent formula is compared to the two SIR phase choices. The SIR phase which is closest

to that calculated by the tangent formula is chosen as the phase which is most probably correct (henceforth called the predicted phase). The probability that a predicted phase is correct should vary between 1.0 and 0.5, since a random selection of one of the two possible phases for each reflection should result in a phase set with 50% of the phases chosen correctly.

As was shown in Part II of the previous Thesis, the ability of the tangent formula to correctly calculate the phase of a reflection is a function of the $|E|$ of the reflection concerned, the amount of error in the initial phase set used by the tangent formula and the resolution to which the data has been collected. The accuracy of the tangent formula increases with increasing $|E|$ and with higher resolution (more terms included in the tangent formula calculations, and hence better statistics) and decreases with increasing error in the initial phase data. Thus, in an application of this method to a specific protein data set, there will exist a certain minimum value of $|E|$ (henceforth called the cutoff E) at which the errors in the phases calculated by the tangent formula will become sufficiently large that the phase choice in step two of the above procedure will become essentially random. The probability of correctly predicting the true phase for reflections at or below the cutoff E should then be 50%. In order to evaluate the results of this method accurately in an application to a real protein data set, the relationship between the probability of a correct phase choice, the $|E|$ of the reflection being determined, the resolution and the amount of error in the initial phase

data must be determined.

One of the parameters in the above relationship, the amount of error in the initial phase data, will be constant for any protein data set and need not be considered. The mean error in a set of SIR centroid phases will be approximately 45°, regardless of the protein considered or the resolution to which the data is collected, and the mean error in the set of phases randomly selected from the two SIR phase choices will be about 90°. The centroid phases are clearly the better of the two sets to use in this method.

The effect of resolution, or the number of terms included in the analysis, is then the only variable parameter which can affect the $|E|$ vs. probability of a correct phase choice curve for a protein SIR data set. This relationship can be determined empirically by the application of the tangent formula phasing procedure to data which have been calculated at a number of resolutions from a theoretical test structure, such as modelglobin in Part II of the preceeding Thesis. This empirical relationship, based on the errorless data calculated from the test structure, can also be applied directly to the data obtained from a real SIR structure determination, as can be seen from the following argument:

In a SIR phase determination, the centroid phase has a value of either $+\Phi_H$ or $-\Phi_H$, depending on the relative values of $|\vec{F}_H|$, $|\vec{F}_P|$ and $|\vec{F}_{PH}|$. The two SIR phase choices are symmetrically situated about the centroid phase angle as in Fig. 1. Except in the case when the two SIR phase angles are close to 90° from the centroid phase angle, a

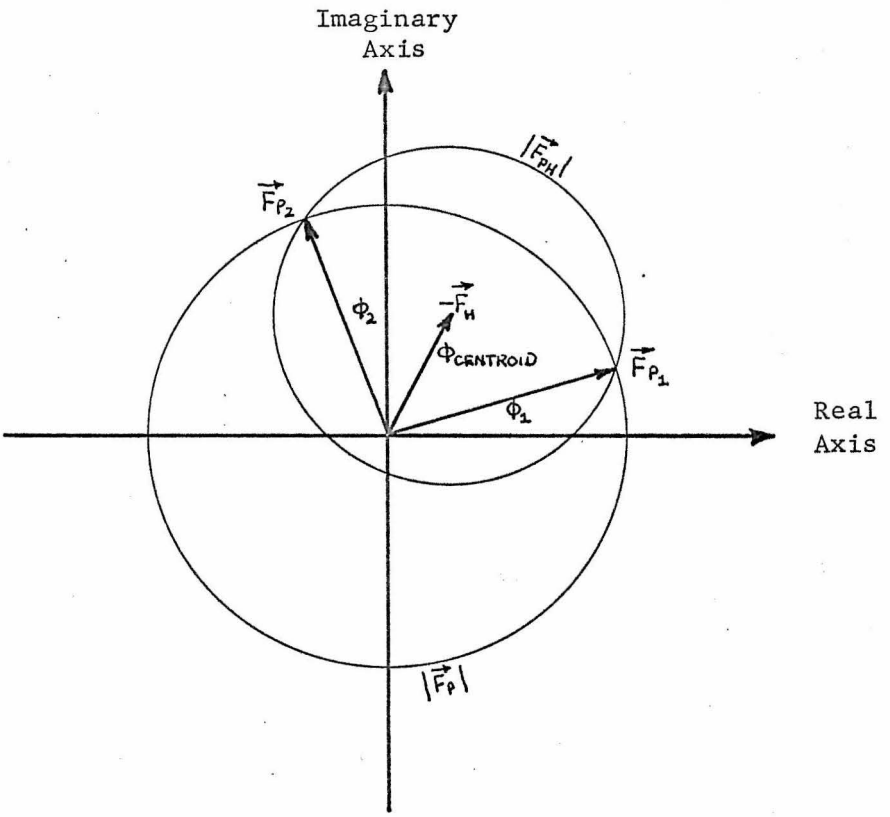

Imaginary
Axis
Real
Axis
$|\vec{F}_{PH}|$
$\vec{F}_{P_2}$
$\Phi_2$
$-\vec{F}_H$
$\Phi_{CENTROID}$
$\vec{F}_{P_1}$
$\Phi_1$
$|\vec{F}_P|$

FIGURE 1

fairly large error in either $|\vec{F}_P|$ or $|\vec{F}_{PH}|$ is necessary to change the centroid phase angle from $+\Phi_H$ to $-\Phi_H$, or vice versa. Thus, the centroid phase set from a real protein SIR structure analysis should be very similar to that calculated from the corresponding test structure. This means that the phases calculated from the tangent formula in step one of the phasing procedure will be about the same in both cases. Figure 2 illustrates this point in the argument. If the calculated phase predicts the correct phase in the theoretical case, the same calculated phase will predict the phase which is closest to the correct phase in the real case, and vice versa. Thus, the probability of choosing the more correct phase from the two possible SIR phase choices is the same in either the real or the theoretical case.

The probability curve above represents the probability that the better of two phases has been chosen, not the probability that the chosen phase is the correct phase. This second probability can be obtained from the original SIR phasing procedure. As can be seen in Fig. 3, the phases of reflections for which the phase circles intersect each other perpendicularly are not highly dependent upon errors in the intensity data, while the opposite is true when the phase circles intersect tangentially. Thus, the accuracy of the two SIR phase angles can be represented by a function of the angle at which the phase circles intersect. The total probability that the chosen phase angle is the correct phase angle (FM 1) can then be taken as some combination of this function and the probability curve above.

In addition to the angle of intersection of the phase circles,

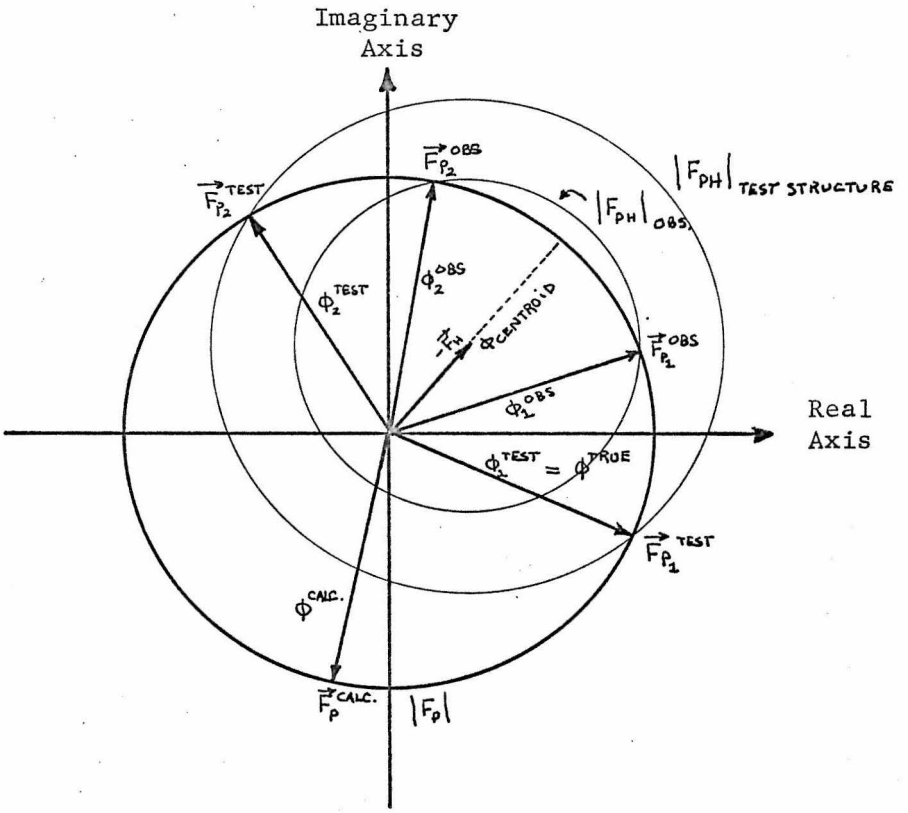

Imaginary
Axis
Real
Axis
$\vec{F}_{P_2}^{OBS}$
$\vec{F}_{P_2}^{TEST}$
$|F_{PH}|_{TEST\ STRUCTURE}$
$|F_{PH}|_{OBS}$
$\phi_2^{TEST}$
$\phi_2^{OBS}$
$\vec{F}_H$
$\phi_{CENTROID}$
$\vec{F}_{P_1}^{OBS}$
$\phi_1^{OBS}$
$\phi_1^{TEST} = \phi^{TRUE}$
$\vec{F}_{P_1}^{TEST}$
$\phi^{CALC.}$
$\vec{F}_P^{CALC.}$
$|F_P|$

FIGURE 2

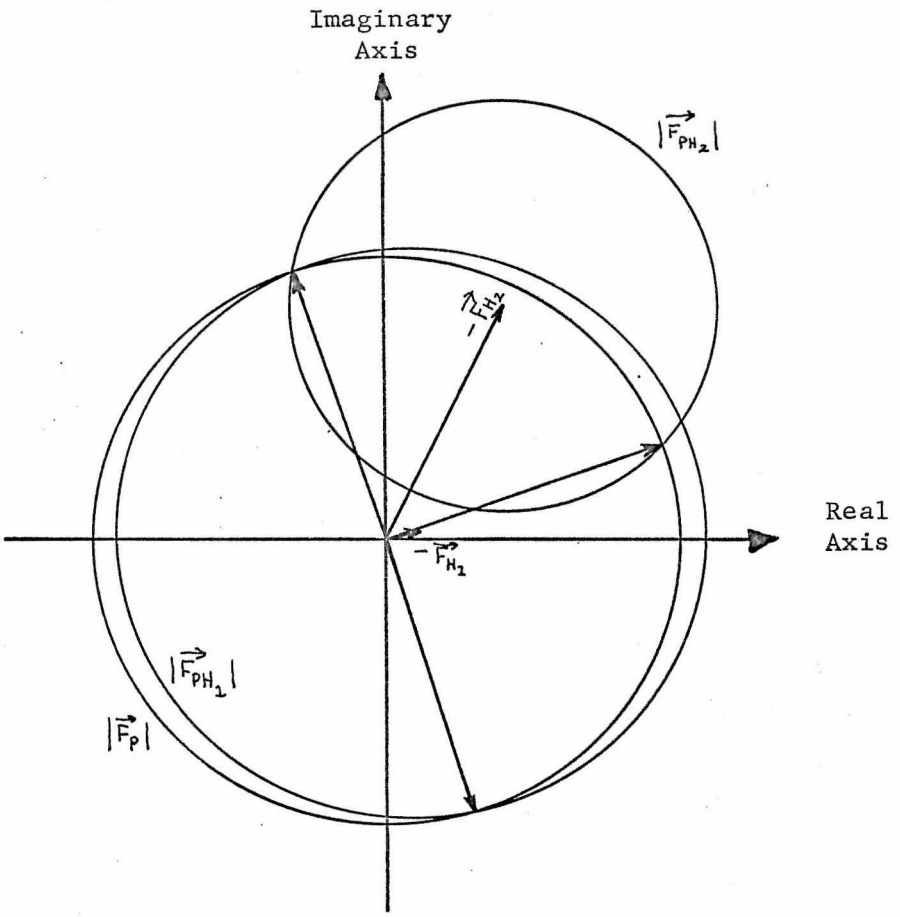



FIGURE 3

The phases from derivative 1 are highly sensitive to changes in $|\vec{F}_{PH_1}|$. The phases from derivative 2 are insensitive to changes in $|\vec{F}_{PH_2}|$.

the relative sizes of the native and derivative phase circles will affect the accuracy of the predicted phase angle. If, as in Fig. 4, the two phase circles differ greatly in size, both of the SIR phase choices, as well as the centroid phase, will be very close to the correct phase. Thus, regardless of the phase angle chosen by the tangent formula phasing procedure, the phase determination will be inherently accurate. This accuracy can be represented by a function of the difference in diameters of the two phase circles (FM 2). When the inherent accuracy of the original SIR phase determination (FM 2) is higher than the previous combined probability function (FM 1), the centroid phase angle should be chosen in place of the angle chosen by the tangent formula phasing procedure in order to obtain the most probable set of phases.

In conclusion, the tangent formula seems to be a very good method for breaking the phase ambiguity in SIR phase data. In order to determine the accuracy of this method as a phase choosing procedure for a real protein SIR data set, it is sufficient to determine its accuracy for a theoretical data set containing the same number of independent reflections. The resulting, theoretical probability curve can be combined with the probability information derived from the original SIR phasing procedure to produce a weighting factor, or figure of merit, for each reflection. This information can then be used to calculate a Fourier map containing the maximum amount of useful structural information.

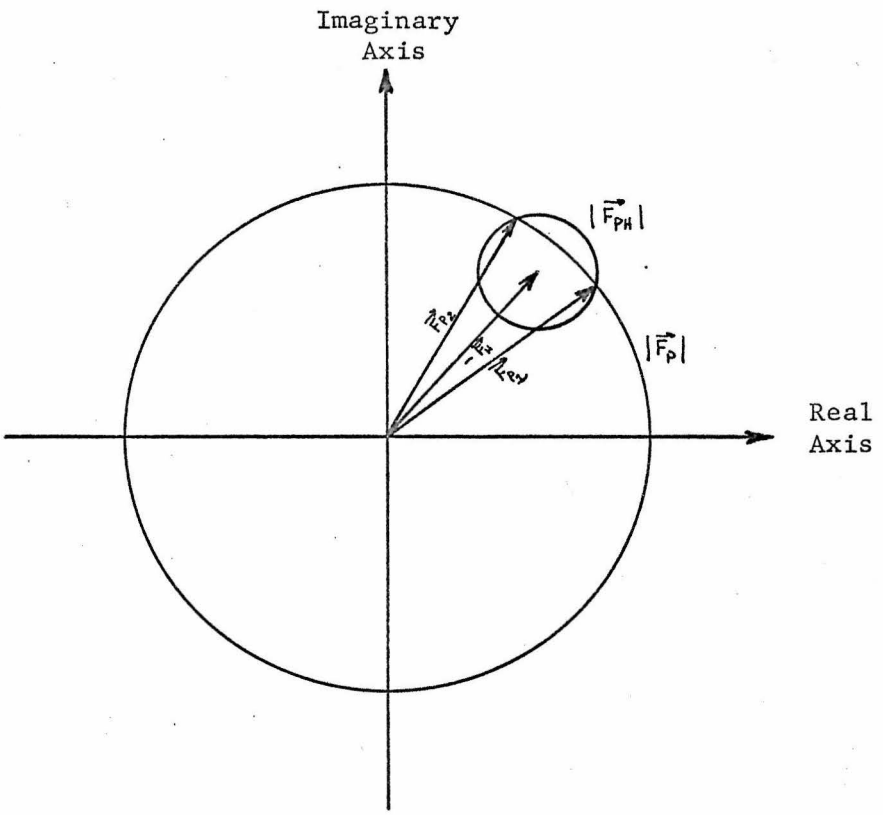



FIGURE 4

## PROPOSITION V

## A METHOD FOR REFINING HIGH RESOLUTION PROTEIN DATA SETS

The refinement of the atomic parameters of a protein structure by the conventional least squares technique is difficult because of the low ratio of observable data to the unknown structural parameters and due to the inherent error in the protein diffraction data. In addition, the necessary set of trial parameters may be difficult to obtain from the final Fourier map with sufficient accuracy to allow the refinement to proceed to the lowest minimum. In order to be able to refine a protein structure accurately then, information other than the intensity data must be included in the refinement procedure.

One possible source for this additional information is the phase information originally obtained from the multiple isomorphous replacement phase analysis. Both the intensities and the multiple isomorphous replacement phases are obtained independently from any structural model of the protein molecule, and thus both are equally well suited to be used as "observed parameters" in the least squares refinement procedure. The multiple isomorphous replacement phases are not independent of the intensity information, however. If the measured intensity of a reflection is subject to a large error, the phase calculated for that reflection will also be in error. This correlation can be removed in practice by including a large number of derivatives in the phase analysis.

As an alternate to a large number of derivatives, the correlation

between the phase information and the intensity information can be removed by tangent formula refinement. In this procedure, the random errors in the phase data which are correlated to the random errors in the intensity data are effectively removed by the averaging process of the tangent formula.

Since the accuracy of the tangent formula in predicting the phase angle of a reflection in a given data set is a function of the magnitude of the normalized structure factor for the reflection, $|E|$, the mean error in the refined phases will rise as reflections of lower $|E|$ are considered and at some minimal value of $|E|$ will become greater than the mean error in the unrefined phases. Below this point, tangent formula refinement cannot be used since it will increase, rather than decrease, the error in the phases.

In order to refine a structure using both intensity and phase information, the following error factor must be minimized:

$$R = \sum_{hkl} w(hkl) \left|\vec{F}_o(hkl) - \vec{F}_c(hkl)\right|^2 \qquad (1)$$

where, $\vec{F}_o(h,k,l)$ represents the observed structure factor magnitude and refined phase, $\vec{F}_c(h,k,l)$ represents the structure factor magnitude and phase angle calculated from the assumed structure and $w(h,k,l)$ represents a weighting factor.

If $\sigma_F(hkl)$ is the standard deviation in $|\vec{F}_o(hkl)|$ and if $\sigma_\Phi(hkl)$ is the standard deviation in $\Phi_o(hkl)$, the probability of finding the specific vector, $\vec{d}(hkl) = \vec{F}_o(hkl) - \vec{F}_c(hkl)$ is (leaving off the

indicies and representing the magnitude of a vector by its symbol without the vector sign):

$$P(\vec{d}) = k \cdot e^{-\frac{1}{2}\left[\frac{d^2}{\sigma_F^2 \cos^2\Phi + \sigma_\Phi^2 F_o^2 \sin^2\Phi}\right]} \tag{2}$$

where $\Phi$ is defined in Fig. 1 and k is a constant normalizing the integrated probability to one. The specific form of Eq. 1 to be normalized is then:

$$R = \sum_{hkl} w(hkl)\, d^2(hkl) \quad , \tag{3}$$

where, $w = \frac{1}{2}(\sigma_F^2 \cos^2\Phi + \sigma_\Phi^2 F_o^2 \sin^2\Phi)$ .

If the weighting factors are considered as constants with respect to the atomic parameters (but are recalculated at the beginning of each refinement cycle) the set of normalized equations to be solved are:

$$\sum_{j=1}^{N} \frac{\partial R}{\partial \xi_j} \frac{\partial R}{\partial \xi_k} \Delta\xi_j = -\frac{\partial R}{\partial \xi_k} \quad ; \quad k = 1, \ldots, N \, , \tag{4}$$

where j and k vary over the atoms and $\xi_j$ varies over the atomic parameters for the j'th atom. Equation 4 can be expanded:

$$\frac{\partial R}{\partial \xi_j} = 2\sum_{hkl} w(hkl) d(hkl) \frac{\partial d(hkl)}{\partial \xi_j} \quad . \tag{5}$$

Dropping the indicies,

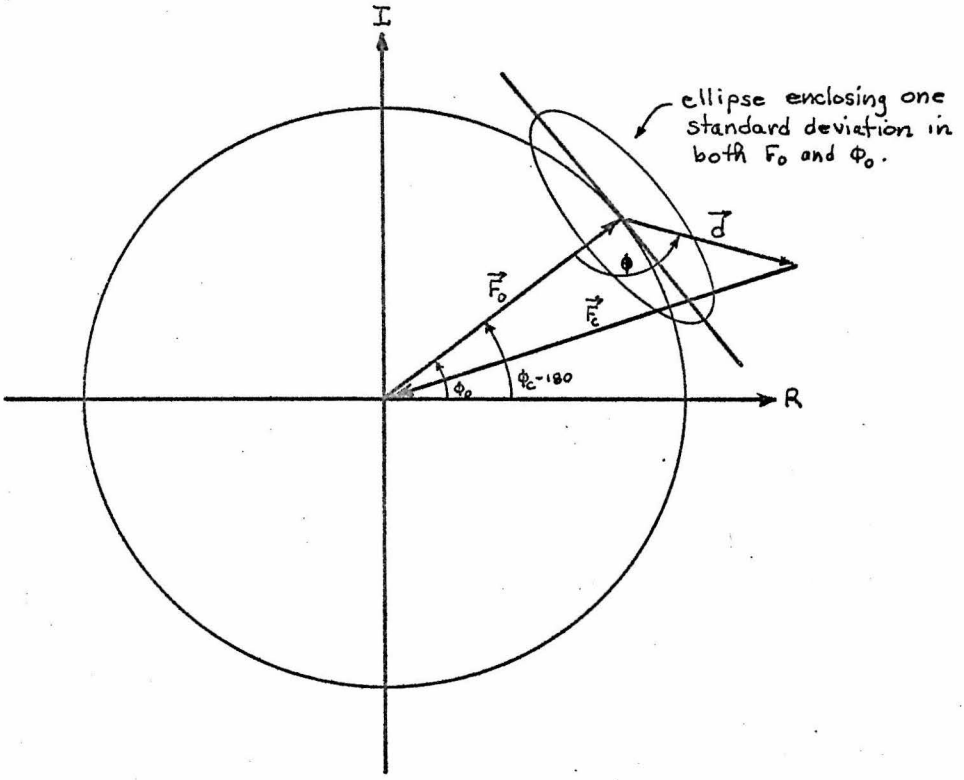



FIGURE 1

$$d^2 = F_o^2 + F_c^2 - 2F_oF_c \cos(\Phi_o - \Phi_c)$$

$$\cos\Phi = \frac{F_o + F_c \cos(\Phi_o - \Phi_c)}{d}$$

$$d = \left[ (A_o - A_c)^2 + (B_o - B_c)^2 \right]^{1/2} \tag{6}$$

$$d \frac{\partial d}{\partial \xi_j} = - \left[ (A_o - A_c) \frac{\partial A_c}{\partial \xi_j} + (B_o - B_c) \frac{\partial B_c}{\partial \xi_j} \right] \tag{7}$$

$$\frac{\partial R}{\partial \xi_j} = -2 \sum_{hkl} w \left[ (A_o - A_c) \frac{\partial A_c}{\partial \xi_j} + (B_o - B_c) \frac{\partial B_c}{\partial \xi_j} \right] \tag{8}$$

For ease of computation, the angle $\Phi(hkl)$ contained in $w(hkl)$ can be redefined:

$$\cos\Phi = \frac{F_o^2 + A_o A_c + B_o B_c}{F_o \left[ (A_o - A_c)^2 + (B_o - B_c)^2 \right]^{1/2}} \tag{9}$$