Theor Chim Acta (1996) 93: 211–233

Theoretica Chimica Acta

Generalized doubly symbolic formulation for integral-driven direct configuration interaction method

Yuji Mochizuki[1], Naoki Nishi[2], Yukio Hirahara[2], Toshikazu Takada[1]

[1] Department of Exploratory Research, Fundamental Research Laboratories, NEC Corporation, Miyukigaoka 34, Tsukuba 305, Ibaraki, Japan; e-mail: mochi@exp.cl.nec.co.jp

[2] Department of Computer System Research, C&C Research Laboratories, NEC Corporation, Miyazkidai 4-1-1, Miyamaeku, Kawasaki 216, Kanagawa, Japan

[3] Network Integration Division, NEC Informatec Systems Corporation, Miyukigaoka 34, Tsukuba 305, Ibaraki, Japan

Received July 3, 1995/Accepted October 10, 1995

Summary. A direct configuration interaction (CI) scheme using the generalized double symbolism both for the external space and for the internal space is proposed in an integral-driven context. The reason why the double symbolism is used in the present formulation is that the main target is in investigating large molecular systems. The integrals, configuration state functions, and energy expressions are systematically classified in terms of the orbital labels and their mutual relations. Various types of CI wavefunctions can be set up flexibly. The resulting structure of integral processings in the sigma vector construction is complicated. The number of unique loop types for two-electron integrals is 1325. Because the parallel architecture is gaining importance in the recent computational platforms, the parallelism is also addressed.

Key words: Configuration interaction – Integral processing – Symbolic expression – Parallelism

1 Introduction

In molecular orbital (MO) calculations, the configuration interaction (CI) method has demonstrated pronounced abilities to treat both the ground and excited states of molecular systems in balanced fashion. Unless otherwise specified, the abbreviation "CI" will hereafter denote the multireference singly and doubly excited version, MRSDCI. The direct CI method that was originated by Roos and Siegbahn [1, 2] should be especially powerful to handle the long expansions. In the direct CI calculations, the so-called sigma vectors are constructed from the molecular integrals and the energy expressions directly in every iteration, where integral classifications are essential. Because of this feature, direct CI does not suffer from the storage and retrieval problems associated with the bulky Hamiltonian matrix elements that are, in contrast, explicitly constructed in the conventional CI scheme. Note that extensive developments of the direct CI methods have been summarized in several monographs (e.g. Refs. [3–6]).

The majority of existing direct CI programs are based on group theories in which the *internal* MOs usually consist both of the active MOs that associate with the open shells or the reference sets to properly describe near-degeneracy effects

and of the internal MOs that are doubly occupied in the reference. The *internal* MOs are treated by the appropriate representation tableau arrays in the program codes, and only the external MO space (unoccupied in the reference) is treated symbolically. In the symbolic CI proposed by Liu and Yoshimine [7], too, only the external symbolism is used. This is because these methods are oriented toward quantitative calculations for the properties of relatively small molecules by incorporating electron correlations that should be flexibly described by the large external space. Definitely, for small molecular systems, today's state-of-the-art direct CI calculations using extensive basis sets with higher polarization functions guarantee satisfactory "chemical accuracy".

In this paper, we present the integral-driven direct CI formulation with a generalized double symbolism. The formulation is oriented toward the qualitative investigation of rather large molecular systems, such as biochemical molecules like drugs, photoactive-dyes and enzyme models, with the standard basis sets. It should be noted that the properties of such systems at the low-lying excited states have been left as fields to be further studied. For example, the chlorophyll dimer is known as the "antenna system" of photosynthesis in green plants, but the characteristics of its excited states, which are of crucial importance to photon capture, are not understood well. The number of active MOs is expected to be still small even for large molecules, but the internal space can be comparable to that of the external space. It is apparent that sizes of the tableau array for the *internal* space and numbers of externally symbolized expressions can become prohibitive and overly redundant when the internal space becomes large. Furthermore, the flexibility with which various types of CI wavefunctions can be constructed is restricted. These difficulties are overcome here by a new scheme in which the expressions are doubly symbolized for both the internal and external spaces. Multi-indexed quantities, such as the integrals and the configuration state functions (CSFs), are systematically classified according to the MO indices and their relations. The present formulation is, of course, based on the extensive works published for the direct CI methods [3–6]. Especially, the key concept "symbolism" depends on Liu and Yoshimine's symbolic CI [7] and also on Saunders and van Lenthe's model-space CI [8]. The basic method for integral classifications owes itself to the pioneering works of Roos and Siegbahn [1, 2]. However, the present CI scheme extends those ideas much more exhaustively because it uses a generalized double symbolism. In fact, for integral processings in the sigma vector construction, the number of unique loop types having respective addressing scheme turns out to be more than thousand. We should describe in detail the handling of symbolic expressions and the control of integral-driven sigma vector loops.

Recently, the parallelization of calculations has been recognized as a promising way to achieve high performance. Most of the current supercomputers are based on the parallel architecture and the workstations (WSs) can now be used in the cluster form through the networks. Thus, we address the parallel applicability. The present formulation is found to be straightforwardly parallelizable due to the integral-driven nature.

2 Direct CI with generalized doubly symbolic energy expressions

2.1 Classification of orbital space

First, the orbital classification should be made explicitly. In the present paper, the whole MO space is divided into five subspaces:

1. Frozen core: always kept doubly occupied or not correlated.
2. Internal: doubly occupied in the reference configurations.
3. Active: used to describe the reference configurations.
4. External: unoccupied in the reference configurations.
5. Frozen external: always kept unoccupied.

Thus, the correlating orbital space consists of the internals, actives, and externals, as mentioned in Sect. 1. The MO labels for these three subspaces are assigned to be $\{i, j, \ldots\}$ for the internals, $\{x, y, \ldots\}$ for the actives, and $\{a, b, \ldots\}$ for the externals. Note that $\{p, q, \ldots\}$ are used as generic indices. The two frozen subspaces need not be concerned with the CI calculation, except for the energy contributions from the frozen core MOs through the effective Fock operators (refer to Sects. 2.3 and 2.4). Further classification according to molecular symmetry is not used in the present paper (although it is potentially applicable) because the symmetry of most large systems, especially for biochemical molecules, is C_1 or "no-symmetry".

2.2 Doubly symbolic CSF set

The CI wavefunctions are described just by the linear expansion of CSFs

$$\Phi^{(R)} = \sum_I T_I^{(R)} \Psi_I, \tag{1}$$

where the CSF set $\{\Psi_I\}$ consists of the reference CSFs, the singly excited CSFs, and the doubly excited CSFs. The CI vectors $\boldsymbol{T}$ are determined as the eigenvectors to diagonalize the Hamiltonian matrix and the energies are obtained as the eigenvalues. The superscript R in Eq. (1) specifies the order of states or simply the CI vector number. A certain CSF is specified by the spatial orbital configuration and the spin couplings of associated open shells and is actually described by the properly grouped Slater determinants. Presently, the excitation is defined not by the spin orbital but by the spatial orbital. In other words, the spin flippings are not regarded as excitations.

A certain orbital configuration is segmented into active orbital part C and the remaining parts of internal and external orbitals that are treated symbolically. Due to combinations of internal and external parts, the total 16 symbolic CSF types are defined as shown in Table 1, and the entire CSF set in Eq. (1) is classified by these doubly symbolic types. This table also contains the occupations in each MO subspace and other possible classifications such as the grand types and the excitation patterns. The grand types are distinguished according to how many electrons move among the subspaces. Caution must be taken, because the symbolic types 1–4 $\{\Psi_0, \Psi_i, \Psi_a, \Psi_{i,a}\}$ (or the grand types 1–4 $\{\Psi_{[V]}, \Psi_{[II]}, \Psi_{[X]}, \Psi_{[IX]}\}$) contain both singly and doubly excited patterns. For example, the Ψ_0 (or $\Psi_{[V]}$) consists of not only the reference CSFs but also the singly ($x \rightarrow y$) and doubly ($xy \rightarrow zw$) excited CSFs associated with only the active MO subspace. The flexibility for setting up orbital configurations could be the most distinctive merit of the direct CI formulation proposed here. Note that the choice of the reference configurations is free or is not restricted to be the complete set within the active MOs. We would exemplify that the grand types $\{\Psi_{[V]}, \Psi_{[I]}, \Psi_{[X]}, \Psi_{[II]}, \Psi_{[IX]}, \Psi_{[IIX]}\}$ are used in the polarization CI (POLCI) and the simpler MRSCI omits the double excitations from this set.

Table 1. CSF classifications

No.	Symbolic type[a]	No. of electrons[b]	Grand class	Excitation pattern[c]
1	Ψ_0	$N_{[I]}/N_{[A]}/0$	$\Psi_{[V]}$	{Ref., $x \rightarrow y$, $xy \rightarrow zw$}
2	Ψ_i	$N_{[I]} - 1/N_{[A]} + 1/0$	$\Psi_{[I]}$	{$i \rightarrow x$, $ix \rightarrow yz$}
3	Ψ_a	$N_{[I]}/N_{[A]} - 1/1$	$\Psi_{[X]}$	{$x \rightarrow a$, $xy \rightarrow za$}
4	$\Psi_{i,a}$	$N_{[I]} - 1/N_{[A]}/1$	$\Psi_{[IX]}$	{$i \rightarrow a$, $ix \rightarrow ya$}
5	Ψ_{i^2}	$N_{[I]} - 2/N_{[A]} + 2/0$	$\Psi_{[II]}$	{$ij \rightarrow xy$}
6	Ψ_{ij}			
7	Ψ_{a^2}	$N_{[I]}/N_{[A]} - 2/2$	$\Psi_{[XX]}$	$xy \rightarrow ab$
8	Ψ_{ab}			
9	$\Psi_{i^2,a}$	$N_{[I]} - 2/N_{[A]} + 1/1$	$\Psi_{[IIX]}$	$ij \rightarrow xa$
10	$\Psi_{ij,a}$			
11	Ψ_{i,a^2}	$N_{[I]} - 1/N_{[A]} - 1/2$	$\Psi_{[IXX]}$	$ix \rightarrow ab$
12	$\Psi_{i,ab}$			
13	Ψ_{i^2,a^2}	$N_{[I]} - 2/N_{[A]}/2$	$\Psi_{[IIXX]}$	$ij \rightarrow ab$
14	$\Psi_{i^2,ab}$			
15	Ψ_{ij,a^2}			
16	$\Psi_{ij,ab}$			

[a] The superscript of "2" for i and a in the CSF label indicates that there are two electrons in the orbital. The relation $i > j$ is satisfied if label "ij" exists (similarly, $a > b$ for the "ab" label)

[b] Numbers of electrons for internals ($N_{[I]}$), actives ($N_{[A]}$), and externals

[c] Excitation patterns are indicated in the symbolic manner

Arbitrary schemes can be used for the spin coupling M for the open-shell parts of the given total orbital configuration, and this is the second merit of the present CI scheme. Namely, both the completely genealogical type and the restricted types are available [9]. The genealogical coupling scheme that is usually employed in group-theory-based programs tends to make the CI length explosive, especially for the case that excited configurations have many open shells with low spins. However, for such a case, the first-order interacting space restriction is useful to dramatically reduce the total CI expansion length without a serious loss of energy [10, 11]. The present flexibility for spin couplings results from the fact that the energy expressions are evaluated by the simple determinant-based particle–hole (PH) method [12–14], as will be discussed in Sect. 2.4. As a whole, both the various classifications of CSF types listed in Table I and the free spin coupling types enable the desired CI wavefunctions to be set up even for large molecular systems.

The CSF address in the vector is defined as follows. The canonical address is used for the indices of internal and external MOs, and the ternary numbers are used to sort the active configuration patterns for each symbolic CSF type. In the example case of five electrons in four active orbitals, the two patterns of "1112" and "2012" are respectively ternary-numbered by 41 and 59, where the active MO index runs from right to left. The addressing for the most complicated type 16 CSF, $\Psi_{ij,ab}(C, M)$, is given by the form [2]

$$\Im(C) - 1 + N_M\{N_{\text{xpair}}[(i-1)(i-2)/2 + j - 1] + (a-1)(a-2)/2 + b - 1\} + M, \tag{2}$$

where $\Im(C)$ is the starting address for the sorted active orbital configuration C, N_M is the number of linearly independent spin couplings for the total orbital configura-

tion $(ij)C(ab)$, and N_{xpair} is the number of possible external orbital pairs. Running ranges for a and b are respectively $(2, N_{\text{ext.}})$ and $(1, N_{\text{ext.}} - 1)$ with the condition of $a > b$. The internal MO indices i and j run similarly. The other symbolic CSF types can be addressed more simply.

In this subsection, the entire CSF set in Eq. (1) has been segmented by introducing the 16 types of symbolic CSFs as

$$\Phi^{(R)} = \sum_{\text{type}}^{16} \left(\sum_{I_{(\text{type})}} T^{(R)}_{I_{(\text{type})}} \Psi_{I_{(\text{type})}} \right). \tag{3}$$

As illustrated in Eq. (2), the second level of segmentation is apparently due to the active orbital configuration C. The segmentation due to C will be crucial in considering the present way of sigma vector construction with the doubly symbolic expressions.

2.3 Hamiltonian matrix elements and classification of integrals

The Hamiltonian matrix elements

$$H_{IJ} = \langle \Psi_I | \hat{H} | \Psi_J \rangle \tag{4}$$

are constructed from the molecular integrals $\boldsymbol{O}$ and the coupling coefficients $\boldsymbol{Q}$ by the formal summation

$$H_{IJ} = \sum_X O_{L_X^{IJ}} Q_X^{IJ}. \tag{5}$$

The suffix X whose length depends on the I/J pair runs over the non-zero contributions. Integrals $\boldsymbol{O}$ consist both of the one-electron integrals (OEIs) $\{h_{pq}\}$ and of the two-electron integrals (TEIs) $\{g_{pq,rs}\}$ (in the charge-cloud notation) and their contributions are specified by the pointer lists $\boldsymbol{L}$. The quantities

$$\{\boldsymbol{L}, \boldsymbol{Q}\} \tag{6}$$

correspond to the energy expressions that will be the central issue in the next subsection.

The definition for the sigma vectors $\boldsymbol{Z}$ is written in matrix form as

$$\boldsymbol{Z} = \boldsymbol{HT}. \tag{7}$$

From Eq. (5), an integral contribution is accumulated directly to the sigma vector element for bra Ψ_I and ket Ψ_J by the pair of arithmetic operations

$$\text{for bra:} \quad Z_I^{(R)} \leftarrow Z_I^{(R)} + O_{L_X^{IJ}} * Q_X^{IJ} * T_J^{(R)}, \tag{8}$$

$$\text{for ket:} \quad Z_J^{(R)} \leftarrow Z_J^{(R)} + O_{L_X^{IJ}} * Q_X^{IJ} * T_I^{(R)}. \tag{9}$$

The sigma vector construction is repeated until the iterative diagonalization procedure converges. Needless to say, the processings of Eqs. (8) and (9) completely dominate the whole CPU time as the kernel of direct CI calculations and should be performed in an efficient manner. One way of efficient processings can be based on the integral-driven context in which all possible I/J pairs associated with the given integral are treated in a group. However, if one uses the integral-driven scheme, the integrals must be classified systematically, as originally pointed out by Roos and Siegbahn [1, 2].

Prior to discussing the classifications, the order of the integrals should be addressed. Unless otherwise noted, the simple "integral" hereafter means the TEI, which is much more laborious to handle than is the OEI, and the processings associated with TEIs will be stressed. It is postulated here that the $\{g_{pq,rs}\}$ are canonically generated and ordered. The canonical indices have the generic form

$$[pqrs] = [pq]([pq] - 1)/2 + [rs], \tag{10}$$

where the $[pq]$ and $[rs]$ are respectively given by

$$[pq] = p(p-1)/2 + q, \qquad [rs] = r(r-1)/2 + s, \tag{11}$$

with the condition of

$$[pq] \geqslant [rs], \quad p \geqslant q, \; r \geqslant s. \tag{12}$$

In the above three equations, the square brackets are used to identify that the indices are packed. Note that the TEI list may be obtained by the standard transformation method from the atomic orbital (AO) integral list. In the direct CI programs with the external symbolism, reorderings for the canonical integral list are often carried out before the sigma vector construction starts. However, the special integral reordering is not considered here, because such a task can be demanding from the viewpoint of data management and retrieval, especially for the calculations of large molecular systems (alternatively, the symbolic energy expressions are reordered, as will be seen in the next subsection). Another reason for the avoidance of reorderings is that the sparsity of the integrals needs to be used. That is, only the contributing integrals within a certain threshold are stored on a file with some index lists and are processed in the sigma vector construction. This is referred to as "prescreening". In fact, as will be seen in Sect. 2.5, our integral-driven method feasibly incorporates the sparsity, and the number of arithmetic operations represented by Eqs. (8) and (9) can effectively be reduced.

Consider now the classifications of the integrals. Table 2 summarizes the numbers of types and subtypes, and the ordered contents, where the total number of subtypes is as many as 94 due to the double symbolism. The integral types are classified according to attributes of the index quartet, and the subtypes are distinguished by their mutual relations. All the subtypes are treated by different integral processing loops. For example, the $g_{xi,yj}$ and $g_{xj,yi}$ integrals (type 5/subtypes 5 and 4, respectively) are processed separately.

If the integrals are canonically stored on a file, the single record associated with the given pq pair generally contains various integrals due to the variations of rs. Table 3 illustrates an example cases in which pq is given by the external orbital pair ab. One finds the total 22 subtypes in this table. As seen in Table 3, if the integral has the active MO labels, this part specifies the supplementary "subaddress" besides the canonical address. The subaddress of the integral will play an essential role in treating the internally and externally symbolic expressions (refer to the next subsection).

As already mentioned, the present paper uses the PH method [12–14] to obtain the doubly symbolic expressions. In the PH method, because any determinants or the resulting CSFs are described only by the "particles" and "holes" relative to the "vacuum determinant" that is not the physically empty state [13], the Hamiltonian matrix elements are given by the vacuum energy (responsible for only the diagonal elements) and the integral contributions are specified only by these particles and holes. In other words, the number of integrals that correlate with a certain H_{IJ} or the length of suffix X in Eq. (5) can be reduced. This should therefore be profitable

Table 2. Integral classifications[a]

Type	No. of subtypes	No. of labels[b]	Ordered contents
1	1	0/4/0	$g_{xy,zw}$
2	1[c]	1/3/0	$g_{xy,zi}$
3	1	0/3/1	$g_{ax,yz}$
4	2	1/2/1	$g_{ai,xy}$, $g_{ax,yi}$
5	5[d]	2/2/0	$g_{xy,ii}$, $g_{xy,ij}$, $g_{xi,yi}$, $g_{xi,yj}$, $g_{xj,yi}$
6	5[d]	0/2/2	$g_{aa,xy}$, $g_{ab,xy}$, $g_{ax,ay}$, $g_{ax,by}$, $g_{ay,bx}$
7	8	3/1/0	$g_{xi,ii}$, $g_{xi,jj}$, $g_{xj,ii}$, $g_{xi,ij}$, $g_{xj,ij}$, $g_{xi,jk}$, $g_{xj,ik}$, $g_{xk,ij}$
8	5	2/1/1	$g_{ax,ii}$, $g_{ax,ij}$, $g_{ai,xi}$, $g_{ai,xj}$, $g_{aj,xi}$
9	5	1/1/2	$g_{aa,xi}$, $g_{ab,xi}$, $g_{ax,ai}$, $g_{ax,bi}$, $g_{ai,bx}$
10	8	0/1/3	$g_{aa,ax}$, $g_{aa,bx}$, $g_{ax,bb}$, $g_{ab,ax}$, $g_{ab,bx}$, $g_{ab,cx}$, $g_{ac,bx}$, $g_{ax,bc}$
11	14	4/0/0	$g_{ii,ii}$, $g_{ii,ij}$, $g_{ii,jj}$, $g_{ii,jk}$, $g_{ij,jj}$, $g_{ij,ij}$, $g_{ij,kk}$, $g_{ik,jj}$, $g_{ij,ik}$, $g_{ij,jk}$, $g_{ik,jk}$, $g_{ij,kl}$, $g_{ik,jl}$, $g_{il,jk}$
12	8	3/0/1	$g_{ai,ii}$, $g_{ai,jj}$, $g_{aj,ii}$, $g_{ai,ij}$, $g_{aj,ij}$, $g_{ai,jk}$, $g_{aj,ik}$, $g_{ak,ij}$
13	9	2/0/2	$g_{aa,ii}$, $g_{ab,ii}$, $g_{aa,ij}$, $g_{ab,ij}$, $g_{ai,ai}$, $g_{ai,bi}$, $g_{ai,aj}$, $g_{ai,bj}$, $g_{aj,bi}$
14	8	1/0/3	$g_{aa,ai}$, $g_{aa,bi}$, $g_{ai,bb}$, $g_{ab,ai}$, $g_{ab,bi}$, $g_{ab,ci}$, $g_{ac,bi}$, $g_{ai,bc}$
15	14	0/0/4	$g_{aa,aa}$, $g_{aa,ab}$, $g_{aa,bb}$, $g_{aa,bc}$, $g_{ab,bb}$, $g_{ab,ab}$, $g_{ab,cc}$, $g_{ac,bb}$, $g_{ab,ac}$, $g_{ab,bc}$, $g_{ac,bc}$, $g_{ab,cd}$, $g_{ac,bd}$, $g_{ad,bc}$

[a] Internal and external indices have respectively the relations of $i > j > k > l$ and $a > b > c > d$

[b] Numbers of labels of internals, actives, and externals in this order

[c] Integrals $g_{xy,zi}$ and $g_{zi,xy}$ actually coexist in the canonical order, but both cases are treated by the single addressing list packing the three active orbital indices (this packed addressing will be referred to as "subaddress" in Table 3). The packing is done here by $N_{\text{acan.}}(z-1)+xy$, where $N_{\text{acan.}}$ is the number of canonical active orbital pairs

[d] For the $g_{xj,yi}$ and $g_{ay,bx}$ integrals, $x > y$ is always satisfied

Table 3. Contents of integral record in a case when pq is given by ab[a]

No.	Integral (type/subtype)	Index area(s)[b]	Subaddress[c]
1	$g_{ab,xy}$ (6/2)		xy
2	$g_{ab,xi}$ (9/2)	i: (1, $N_{\text{int.}}$)	x
3	$g_{ab,ax}$ (10/4)		x
4	$g_{ab,bx}$ (10/5)		x
5	$g_{ab,cx}$ (10/6)	c: (1, $b-1$)	x
6	$g_{ac,bx}$ (10/7)	b: ($c+1$, $a-1$)	x
7	$g_{ab,ii}$ (13/2)	i: (1, $N_{\text{int.}}$)	
8	$g_{ab,ij}$ (13/4)	i: (2, $N_{\text{int.}}$), j: (1, $i-1$)	
9	$g_{ab,ai}$ (14/4)	i: (1, $N_{\text{int.}}$)	
10	$g_{ab,bi}$ (14/5)	i: (1, $N_{\text{int.}}$)	
11	$g_{ab,ci}$ (14/6)	c: (1, $b-1$)	
12	$g_{ac,bi}$ (14/7)	b: ($c+1$, $a-1$)	
13	$g_{ab,bb}$ (15/5)		
14	$g_{ab,ab}$ (15/6)		
15	$g_{ab,cc}$ (15/7)	c: (1, $b-1$)	
16	$g_{ac,bb}$ (15/8)	b: ($c+1$, $a-1$)	
17	$g_{ab,ac}$ (15/9)	c: (1, $b-1$)	
18	$g_{ab,bc}$ (15/10)	c: (1, $b-1$)	
19	$g_{ac,bc}$ (15/11)	b: ($c+1$, $a-1$)	
20	$g_{ab,cd}$ (15/12)	c: (2, $b-1$), d: (1, $c-1$)	
21	$g_{ac,bd}$ (15/13)	b: ($c+1$, $a-1$), d: (1, $c-1$)	
22	$g_{ad,bc}$ (15/14)	b: ($d+2$, $a-1$), c: ($d+1$, $b-1$)	

[a] Actually, because of the prescreening, only the non-zero integrals (within a certain threshold) are stored with an appropriate index list

[b] Possible running index area for the internal and external MO spaces. The paired numbers in parentheses indicate the bottom and top of the area

[c] The subaddress is defined by packing only the active MO indices. See the caption for the type 2 integral in Table 2

for the integral-driven sigma vector construction, especially for molecules having a large internal MO space. Here, the modified OEI elements of the effective Fock-type operator and the vacuum energy are given by

$$f_{pq} = h_{pq} + \sum_{r}^{\text{Vac. Det.}} (2g_{pq,rr} - g_{pr,qr}), \tag{13}$$

$$E_{\text{Vac.}} = \sum_{r} (f_{rr} + h_{rr}), \tag{14}$$

where the summation runs over all the occupied orbitals in the vacuum determinant of the closed-shell type. The number of closed shells is given by the integer part of half the total number of electrons included in the target molecular system. The energy contributions from the frozen core MOs (refer to Sect. 2.1) are included through Eqs. (13) and (14). Once the OEIs having the canonical addresses are generated, these two-indexed quantities are much more simply classified than are those in the four-indexed TEI case. There are six OEI types, and the total number of subtypes is only eight. In summary, the OEI treatment is much easier than the TEI treatment.

2.4 *Doubly symbolic energy expression*

In the previous two subsections, the CSFs and the molecular integrals are classified according to the orbital classes and their relations. The energy expressions should therefore be classified consistently. As mentioned in Sect. 2.2, a certain CSF is characterized by the active orbital configuration C, the configuration pattern due to the internal and external orbitals that are treated symbolically, and the spin coupling M. Thus, for example, the actual Hamiltonian matrix elements for the interaction between type 16 CSFs may be represented as

$$\langle 16|\hat{H}|16\rangle_{IJ} = \langle \Psi_{ij,ab}(C, M)|\hat{H}|\Psi_{kl,cd}(C', M')\rangle, \tag{15}$$

where the elements are specified by a total of 12 indices (ten for the configuration part and two for the spin coupling part.) The minimum number of indices is four for the $\langle 1|\hat{H}|1\rangle$ interaction in which there is (of course) no symbolism. The other CSF interactions fall into the intermediate cases. However, because the electron–electron interaction is at most two-body in nature, common orbitals between the bra and ket must exist in the non-zero $\boldsymbol{H}$ elements. Presently, for the integral-driven sigma vector construction of Eqs. (8) and (9), the expressions should be handled with the internally and externally double symbolism and should therefore be structured carefully. This subsection discusses in detail a systematic method used to obtain such symbolic expressions as compact as possible and to determine the vector addressings for $\boldsymbol{T}$ and $\boldsymbol{Z}$.

First, we consider how to define the symbolic Hamiltonian matrix elements to be explicitly evaluated. Recall that all the configurations are characterized by the active orbital configuration, the maximal two internals, and the maximal two externals. In a certain matrix element, the combination of active orbital configuration C/C' is unique, where the number of electrons involved with each configuration need not be the same. Thus, we have to be concerned with the symbolic combination patterns of internal and/or external orbitals between the bra and ket. When the bra/ket pair of the symbolic CSF types is given, the internal and external orbital combinations in the matrix element are respectively classified by each of primary label types (PLTs), according to the list in Table 4. Taking the $\langle 16|\hat{H}|14\rangle$ interaction of $\langle \Psi_{ij,ab}(C, M)|\hat{H}|\Psi_{k^2,cd}(C', M')\rangle$ as an example, the internal PLT is I (for ij/k^2), and the external PLT is J (for ab/cd). A more specific description, taking account of identities and inequalities between the individual indices, is characterized by the secondary label types (SLTs), separately for the internals and externals. The correspondences between PLT and SLT are listed in Table 5. Note that the PLT J has 13 cases of associated SLTs. Thus, any bra/ket configuration pairs to be evaluated within the symbolic $\boldsymbol{H}$ space are uniquely defined by using combination lists of the symbolic CSF type, PLT/SLT, and active orbital configuration.

The setting up of the symbolic MO space is a simple matter. In the usual CI cases, four symbolic internals and also four symbolic externals are enough for constructing the symbolic $\boldsymbol{H}$. The energy expressions for the symbolic matrix elements are obtained through the determinant-based PH technique that is based just on the second quantization formalism and the Slater rules. There are ten types of elementary expressions for the determinant pairs [14]. An integral address itself is determined by the character of the given symbolic orbital configuration pair, and only the non-zero combination of spin coupling M/M' is kept for the final coupling coefficients for this configuration pair. A pronounced advantage of the simple

Table 4. Primary label types (PLTs)

Case	Bra/ket label pair[a]
A	None, $p/$, $/p$, $p^2/$, $/p^2$, $pq/$, $/pq$
B	p/q
C	p/q^2
D	p^2/q
E	p/qr
F	pq/r
G	p^2/q^2
H	p^2/qr
I	pq/r^2
J	pq/rs

[a] Case values are separately assigned to the internal and external label parts of the symbolic matrix elements (see also the text)

Table 5. Correspondences between **primary label type (PLT)** and **secondary label type (SLT)**

Primary label type	Index relation	Secondary label type
A	None	0
B, G	$p = q$, $p^2 = q^2$	1
	$p > q$, $p^2 > q^2$	2
	$p < q$, $p^2 < q^2$	3
C, D	$p = q^2$, $p^2 = q$	1
	$p > q^2$, $p^2 > q$	2
	$p < q^2$, $p^2 < q$	3
E, H	$p = q$, $p^2 = q$	1
	$p = r$, $p^2 = r$	2
	$p > q$, $p^2 > q$	3
	$q > p > r$, $q > p^2 > r$	4
	$q > r > p$, $q > r > p^2$	5
F, I	$p = r$, $p = r^2$	1
	$q = r$, $q = r^2$	2
	$q > r$, $q > r^2$	3
	$p > r > q$, $p > r^2 > q$	4
	$r > p$, $r^2 > p$	5
J	$(p = r, q = s)$	1
	$(p = r, q > s)$	2
	$(p = r, q < s)$	3
	$(p > r, q = s)$	4
	$(p < r, q = s)$	5
	$(p = s, q < r)$	6
	$(p > s, q = r)$	7
	$q > r$	8
	$p > r > q > s$	9
	$p > r > s > q$	10
	$p < s$	11
	$q < s < p < r$	12
	$(p < r, q > s)$	13

Table 6. Positional pattern types (PPTs) for an integral index t

Case	Bra/ket label pair[a]
0	None
1	t/t, tp/tq
2	pt/qt
3	t^2/t^2
4	$t/$, $tp/$
5	$/t$, $/tp$
6	$pt/$
7	$/pt$
8	$t^2/$
9	$/t^2$
10	pt/t, pt/tq
11	t/pt, tp/qt
12	t^2/t, t^2/tp
13	t/t^2, tp/t^2
14	t^2/pt
15	pt/t^2

[a] If t is an internal orbital index, the bra/ket label pair refers to the internal orbital indices of the matrix element. The case of externals is similar to that of internals

determinant-based evaluations is that arbitrary spin coupling schemes can be used in the CSF set, as mentioned in Sect. 2.2. The speed for methods based on the group theories can be faster than that for the determinant-based method. However, because the size of the expressions is minimized due to the double symbolism, this would not become serious.

The energy expressions are evaluated explicitly for the symbolic $\boldsymbol{H}$ elements. The $\boldsymbol{L}$ part of each of these expressions contains the various types/subtypes of integrals, which have been classified as shown in Table 2. Because this situation is apparently in conflict with the integral-driven processing, the expressions for symbolic elements must be reordered according to the classified integrals or the types/subtypes of integrals in the backward manner.

In the integral-driven context, a typed/subtyped symbolic integral correlates with the list of various symbolic matrix elements. The indices of the integral must be attributed to the bra indices and/or the ket indices of the element, and the common indices between the bra and ket must be found. More specifically, to determine the loop addressings in the sigma vector construction, the position of each of the symbolic integral indices needs to be found in the indices of each symbolic matrix element. For this purpose, the positional pattern types (PPTs) are defined. There are 16 cases of PPTs as listed in Table 6. The PPT of an active orbital index is always zero (because the active orbitals do not appear in the symbolic labels of matrix elements but relate directly to the active orbital configuration part), but the other integral indices are attributed to the symbolic labels without regard to whether they are internal or external. In the actual sigma vector construction, the symbolic internal and external MO indices of integrals are replaced by those of integrals stored canonically on a file and the common orbital

Table 7. Loop characteristics of $\langle 16|\hat{H}|10\rangle$ interaction for $g_{ai,xi}$ (type 8/subtype 3)

No.	Matrix element[a,b]	SLTs[c]	PPTs[d]	Common patterns[b,e]
1	$\langle \Psi_{i\tilde{j},\tilde{b}a}\|\hat{H}\|\Psi_{i\tilde{j},\tilde{b}}\rangle$	1 1	6 1 0 1	down up
2	$\langle \Psi_{i\tilde{j},a\tilde{b}}\|\hat{H}\|\Psi_{i\tilde{j},\tilde{b}}\rangle$	1 2	4 1 0 1	down down
3	$\langle \Psi_{\tilde{j}i,\tilde{b}a}\|\hat{H}\|\Psi_{\tilde{j}i,\tilde{b}}\rangle$	1 1	6 2 0 2	up up
4	$\langle \Psi_{\tilde{j}i,a\tilde{b}}\|\hat{H}\|\Psi_{\tilde{j}i,\tilde{b}}\rangle$	1 2	4 2 0 2	up down

[a] For simplicity, the active orbital configurations and the spin couplings are not shown in the matrix element. The general form of the element is described as $\langle \Psi_{ij,ab}|\hat{H}|\Psi_{kl,c}\rangle$, where the PLTs for the internal and external parts are respectively J and F

[b] The common orbitals between the bra and ket are identified by the tilde. Hereafter, the common orbital index is simply called common

[c] Values for the internal and external label parts

[d] Values for the integral index quartet of "*a i x i*"

[e] Patterns for the internal and external commons. The definitions of pattern are summarized in Table 8

indices run in the loops whose addressing controls depend both on the PPTs and on the bra/ket CSF types. If the integral indices contain the active MOs or the integral has the subaddress, this index part is responsible for the specification of the pair of active orbital configurations C/C' of interacting CSFs, as will be exemplified in the next paragraph. The number of the PPTs to define the loops for TEI processings is obviously four. That is, the PPT quartets are necessary for TEI. Note that the total number of unique loop type for TEI turns out to be 1325. In contrast to the TEI case, the number of loops for OEI is only 144 because of the two-indexed nature using the PPT doublets.

We would like to introduce an example to illustrate the structure of symbolic expressions and the addressing scheme for the vectors. Table 7 shows the loop characteristics of the $\langle 16|\hat{H}|10\rangle$ interaction for the $g_{ai,xi}$ (type 8/subtype 3) integral. One finds that, although the number of the SLT combinations is two ("1 1" and "1 2"), the interaction has a total of four loops due to the four PPT quartets "6 1 0 1", "4 1 0 1", "6 2 0 2", and "4 2 0 2". This is the reason why not the SLT but the PPT must be used to define the loop addressings. As indicated by the associated matrix elements in Table 7, the internal index i has the two PPT cases of "1" by $i\tilde{j}/i\tilde{j}$ and "2" by $\tilde{j}i/\tilde{j}i$, where the tilde identifies the common orbital already referred to simply as the common in the table. Similarly, a has "4" by $a\tilde{b}/\tilde{b}$ and "6" by $\tilde{b}a/\tilde{b}$. The running areas of commons ($\tilde{j}$ and $\tilde{b}$) are restricted by the orbital indices having the non-zero PPTs (i and a, respectively). The pattern names for these commons are also given in Table 7, and the origin of the names is found in Table 8. There are a total of five patterns of commons (the other three patterns will be introduced in the next subsection). The relations of $\tilde{j}i/\tilde{j}i$ and $\tilde{b}a/\tilde{b}$ correspond to the up case, and $i\tilde{j}/i\tilde{j}$ and $a\tilde{b}/\tilde{b}$ represent the down case. Because of the canonical order of the TEI list, only the active orbital index x, that corresponds just to the subaddress, varies within a given single record (that is, the a and i are predetermined in the record). As already noted, the active MO space is handled without any symbolism. In this example, x associates directly with the active orbital configuration pair C/C'. Namely, if the integral has the subaddress due to active MOs, the list of C/C' is determined not only by the type/subtype of integral but also by its subaddress. Note that the (non-zero) coupling coefficients have the

Table 8. Definitions of commons

Name	Running area(s)[a]
up	$\tilde{i}$: (Max[j_{bra}, k_{ket}] + 1, $N_{\text{int.}}$) $\tilde{a}$: (Max[b_{bra}, c_{ket}] + 1, $N_{\text{ext.}}$)
down	$\tilde{i}$: (1, Min[j_{bra}, k_{ket}] − 1) $\tilde{a}$: (1, Min[b_{bra}, c_{ket}] − 1)
middle	$\tilde{i}$: (Min[j_{bra}, k_{ket}] + 1, Max[j_{bra}, k_{ket}] − 1) $\tilde{a}$: (Min[b_{bra}, c_{ket}] + 1, Max[b_{bra}, c_{ket}] − 1)
solo	$\tilde{i}$: (1, $N_{\text{int.}}$) $\tilde{a}$: (1, $N_{\text{ext.}}$)
duo	$\tilde{i}$: (2, $N_{\text{int.}}$), $\tilde{j}$: (1, $N_{\text{int.}} - 1$), $i > \tilde{j}$ $\tilde{a}$: (2, $N_{\text{ext.}}$), $\tilde{b}$: (1, $N_{\text{ext.}} - 1$), $\tilde{a} > \tilde{b}$

[a] The symbol Max[p_{bra}, q_{ket}] specifies the larger of p from the bra and q from the ket, where both p and q are the integral indices having non-zero PPTs. If either p_{bra} or q_{ket} is missing, the existing one is taken. Min[p_{bra}, q_{ket}] works similarly

dependence both on the spin coupling pair M/M' and on C/C'. For the PPT "6 1 0 1" quartet, the internal and external orbital parts of CSF addressings in the loop are given by

$$N_{\text{xpair}}[(i-1)(i-2)/2 + \tilde{j} - 1] + (\tilde{b} - 1)(\tilde{b} - 2)/2 + a - 1 \tag{16}$$

for the bra or type 16 CSF [see also Eq. (2)] and by

$$N_{\text{ext.}}[(i-1)(i-2)/2 + \tilde{j} - 1] + \tilde{b} - 1 \tag{17}$$

for the ket or type 10 CSF, and the ranges of $\tilde{j}$ and $\tilde{b}$ are respectively $(1, i-1)$ of down and $(a+1, N_{\text{ext.}})$ of up. The remaining parts of CSF addressings are specified in the outer loops due to M/M', C/C', and x (as will be seen in Fig. 1). In this way, the contributions from each $g_{ai,xi}$ integral to the sigma vector of the type 16 and 10 CSF blocks should be correctly calculated in each loop for every state R.

As described in this subsection, the reordered structure of doubly symbolic expressions for a certain integral specified by the type/subtype is characterized by the list of CSF interactions, that of PPT values which define the loop addressing schemes through the commons, that of active orbital configuration pairs, and that of spin coupling pairs, in this order of the lists. If the integral has the supplementary subaddress, the active orbital configuration list depends also on the subaddress.

The energy expressions may finally be written on a direct-access type file. For the usual cases having some few reference configurations, the size of final expressions is at most a few megabytes and all the contents can be read into a central memory before the iterative sigma vector construction starts. Thus, I/O operations about the expressions would be avoided during the iteration.

2.5 Integral-driven sigma vector construction

Before discussing the sigma vector construction itself, the variety of loops should be considered. Table 9 lists the numbers of different loops for each of the integrals and

Table 9. Number of loops for integrals and number of associated interactions

Type	Subtype	Loops	Interactions
1	1	16	16
2	1	16	12
3	1	16	12
4	1	16	9
	2	32	18
5	1	16	12
	2	40	12
	3	20	16
	4	44	16
	5	28	16
6	1	16	12
	2	40	12
	3	20	16
	4	44	16
	5	28	16
7	1	4	4
	2	4	4
	3	4	4
	4	8	8
	5	8	8
	6	8	4
	7	8	4
	8	8	4
8	1	16	9
	2	48	12
	3	20	12
	4	28	15
	5	28	15
9	1	16	9
	2	48	12
	3	20	12
	4	28	15
	5	28	15
10	1	4	4
	2	4	4
	3	4	4
	4	8	8
	5	8	8
	6	8	4
	7	8	4
	8	8	4
11	1	4	4
	2	4	4
	3	4	4
	4	4	4
	5	4	4
	6	8	8

Table 9. Continued

Type	Subtype	Loops	Interactions
	7	4	4
	8	4	4
	9	8	8
	10	8	8
	11	8	8
	12	8	4
	13	8	4
	14	8	4
12	1	4	3
	2	4	3
	3	4	3
	4	8	6
	5	8	6
	6	8	3
	7	8	3
	8	8	3
13	1	16	9
	2	24	9
	3	24	9
	4	72	10
	5	17	10
	6	25	10
	7	25	10
	8	37	11
	9	37	11
14	1	4	3
	2	4	3
	3	4	3
	4	8	6
	5	8	6
	6	8	3
	7	8	3
	8	8	3
15	1	4	4
	2	4	4
	3	4	4
	4	4	4
	5	4	4
	6	8	8
	7	4	4
	8	4	4
	9	8	8
	10	8	8
	11	8	8
	12	8	4
	13	8	4
	14	8	4

Table 10. Loop characteristics for $g_{ij,kl}$ (type 11/subtype 12)

No.	Interaction	Matrix element	PPTs	Common pattern
1	$\langle 6\|\hat{H}\|6\rangle$	$\langle \Psi_{ik}\|\hat{H}\|\Psi_{jl}\rangle$	4 5 6 7	
2		$\langle \Psi_{il}\|\hat{H}\|\Psi_{jk}\rangle$	4 5 7 6	
3	$\langle 10\|\hat{H}\|10\rangle$	$\langle \Psi_{ik,\bar{a}}\|\hat{H}\|\Psi_{jl,\bar{a}}\rangle$	4 5 6 7	solo
4		$\langle \Psi_{il,\bar{a}}\|\hat{H}\|\Psi_{jk,\bar{a}}\rangle$	4 5 7 6	solo
5	$\langle 15\|\hat{H}\|15\rangle$	$\langle \Psi_{ik,\bar{a}^2}\|\hat{H}\|\Psi_{jl,\bar{a}^2}\rangle$	4 5 6 7	solo
6		$\langle \Psi_{il,\bar{a}^2}\|\hat{H}\|\Psi_{jk,\bar{a}^2}\rangle$	4 5 7 6	solo
7	$\langle 16\|\hat{H}\|16\rangle$	$\langle \Psi_{ik,\bar{a}\bar{b}}\|\hat{H}\|\Psi_{jl,\bar{a}\bar{b}}\rangle$	4 5 6 7	duo
8		$\langle \Psi_{il,\bar{a}\bar{b}}\|\hat{H}\|\Psi_{jk,\bar{a}\bar{b}}\rangle$	4 5 7 6	duo

the total number of loops is summed up to 1325. In this table, the numbers of associated interactions between the CSF types are also included. Each symbolic CSF interaction has a multiplicity of addressing patterns due to the variety of the PPT quartet, as seen for the $g_{ai,xi}$ case in the last subsection. To introduce the three new patterns of commons, the loops for the $g_{ij,kl}$ (type 11/subtype 12) and $g_{ab,ij}$ (type 13/subtype 4) integrals are taken as examples. The eight cases for $g_{ij,kl}$ loops are listed in Table 10, where the four different interactions are involved. This example shows two patterns of new commons named solo and duo (refer to Table 8). Without the dependence on integral indices, these commons run over the possible space of each of the internal and external MOs in the actual processings (the external space in this example). The largest number of loop addressing patterns is 18 for the $\langle 16|\hat{H}|16\rangle$ interaction of the $g_{ab,ij}$ integral, whose characteristics are listed in Table 11. A restricted common newly appears in this table. The new common is named middle, where "middle" means that both the bottom and top of the index range to be run are restricted by the integral indices. Note that, in these two examples of integrals, the active orbital configurations and spin couplings in the matrix element have the diagonal relations of $C = C'$ and $M = M'$, respectively. Even from the three introduced integrals of $g_{ai,xi}$, $g_{ij,kl}$, and $g_{ab,ij}$, one can imagine the general scheme of addressings for the $\boldsymbol{T}$ and $\boldsymbol{Z}$ and can be seen that the variety of TEI loops is systematic and reasonable although its total number is as many as 1325. Attention should be paid again to the fact that each loop characteristic is uniquely specified not by the commons but by the PPT quartets.

Consider now the layered loop structure of our sigma vector construction and its control flow. Figure 1 schematically illustrates the form of layers. The outermost or first layer corresponds to the specification of pq for the canonical TEI list. Then the records of non-zero integral lists, which have been prescreened with a certain threshold, are read from a file for the given pq. Recall that a single integral record contains the various types/subtypes due to $\{rs\}$ (as shown in Table 3) and each of the types/subtypes is correlated with many interactions between the CSF types (Table 9). Thus, the second level of layers is concerned with the integral types/subtypes included in a record. When the type/subtype of integrals is fixed, the associated symbolic energy expressions are read from a file, where the variations due to the subaddresses must properly be taken into account if the integral indices contain the active orbital label (again, refer to Table 3). The kernel of loops in the

Table 11. Loop characteristics of $\langle 16|\hat{H}|16\rangle$ interaction for $g_{ab,ij}$ (type 13/subtype 4)

No.	Matrix element	PPTs	Common patterns[a]	
1	$\langle \Psi_{\tilde{k}i,\tilde{c}a}\|\hat{H}\|\Psi_{\tilde{k}j,\tilde{c}b}\rangle$	6767	up	up
2	$\langle \Psi_{\tilde{k}i,\tilde{c}b}\|\hat{H}\|\Psi_{\tilde{k}j,\tilde{c}a}\rangle$	7667	up	up
3	$\langle \Psi_{\tilde{k}i,a\tilde{c}}\|\hat{H}\|\Psi_{\tilde{k}j,\tilde{c}b}\rangle$	4767	up	middle
4	$\langle \Psi_{\tilde{k}i,\tilde{c}b}\|\hat{H}\|\Psi_{\tilde{k}j,a\tilde{c}}\rangle$	5667	up	middle
5	$\langle \Psi_{\tilde{k}i,a\tilde{c}}\|\hat{H}\|\Psi_{\tilde{k}j,b\tilde{c}}\rangle$	4567	up	down
6	$\langle \Psi_{\tilde{k}i,b\tilde{c}}\|\hat{H}\|\Psi_{\tilde{k}j,a\tilde{c}}\rangle$	5467	up	down
7	$\langle \Psi_{i\tilde{k},\tilde{c}a}\|\hat{H}\|\Psi_{j\tilde{k},\tilde{c}b}\rangle$	6747	middle	up
8	$\langle \Psi_{i\tilde{k},\tilde{c}b}\|\hat{H}\|\Psi_{j\tilde{k},\tilde{c}a}\rangle$	7647	middle	up
9	$\langle \Psi_{i\tilde{k},a\tilde{c}}\|\hat{H}\|\Psi_{j\tilde{k},\tilde{c}b}\rangle$	4747	middle	middle
10	$\langle \Psi_{i\tilde{k},\tilde{c}b}\|\hat{H}\|\Psi_{j\tilde{k},a\tilde{c}}\rangle$	5647	middle	middle
11	$\langle \Psi_{i\tilde{k},a\tilde{c}}\|\hat{H}\|\Psi_{j\tilde{k},b\tilde{c}}\rangle$	4547	middle	down
12	$\langle \Psi_{i\tilde{k},b\tilde{c}}\|\hat{H}\|\Psi_{j\tilde{k},a\tilde{c}}\rangle$	5447	middle	down
13	$\langle \Psi_{l\tilde{k},\tilde{c}a}\|\hat{H}\|\Psi_{j\tilde{k},\tilde{c}b}\rangle$	6745	down	up
14	$\langle \Psi_{l\tilde{k},\tilde{c}b}\|\hat{H}\|\Psi_{j\tilde{k},\tilde{c}a}\rangle$	7645	down	up
15	$\langle \Psi_{l\tilde{k},a\tilde{c}}\|\hat{H}\|\Psi_{j\tilde{k},\tilde{c}b}\rangle$	4745	down	middle
16	$\langle \Psi_{l\tilde{k},\tilde{c}b}\|\hat{H}\|\Psi_{j\tilde{k},a\tilde{c}}\rangle$	5645	down	middle
17	$\langle \Psi_{l\tilde{k},a\tilde{c}}\|\hat{H}\|\Psi_{j\tilde{k},b\tilde{c}}\rangle$	4545	down	down
18	$\langle \Psi_{l\tilde{k},b\tilde{c}}\|\hat{H}\|\Psi_{j\tilde{k},a\tilde{c}}\rangle$	5445	down	down

[a] Patterns for the internal and external commons

present integral-driven sigma vector construction starts at this point. The expressions first specify the interactions of CSF types and then do the PPT list. Thus, the third and fourth layers correspond respectively to the CSF interactions and the PPT quartets, both of which correlate with the internal and external orbital parts of CSF addressings through the commons having five patterns. The inner multiloop can be in the order of the state number R, rs, the active orbital configuration pair C/C', the spin coupling pair M/M', and the commons. There are two points to be recalled. First, the running range of commons is dependent generally on the pq and rs indices that are pointing to the prescreened integral entries. Second, the loop over M/M' is driven only by the non-zero values of coupling coefficients. As a result, arithmetic operations due to the non-contributive integrals and coupling coefficients are naturally avoided. This feature is notably favorable to executions on the scalar machines such as WSs. The loop over commons is currently placed to be innermost and can be executed with the vectorization on the platform having the vector CPU. In the actual implementation, if one wishes to elongate the vector length for efficiency, one can fuse the loops of commons, M/M', etc. or can alter the loop order, depending upon each of the integral types/subtypes. However, such a discussion on the tunings is somewhat out of the purpose of the present paper. In any case, the bra and ket CSF addresses are calculated, and according to them the integral contributions are taken into account through the pair of vector-addition operations corresponding to Eqs. (8) and (9).

As just noted in the above paragraph, the CI formulation proposed presently incorporates the sparsity due to the integral prescreening into the sigma vector construction. In the result, the meaningless operations are bypassed. The use of the localized MO sets [15] can increase the efficiency of prescreenings. A more drastic way to save the CPU time may be a controlled neglect of certain types/subtypes of

```
Loop over pq
  Read prescreened integrals due to {rs}
  Loop over integral types/subtypes
    Read associated energy expressions
    Loop over CSF interactions
      Loop over PPT quartets
        Loop over R
          Loop over rs
            Loop over C/C'
              Loop over M/M'
                Loop over commons
                  Bra-address ← f(CSF-type, PPT, {p, q, r, s}, C, commons, M)
                  Ket-address ← f'(CSF-type, PPT, {p, q, r, s}, C', commons, M')
                  Z(Bra-address, R) ← Z(Bra-address, R) +
                      O(rs)*Q(rs, C/C', M/M')*T(Ket-address, R)
                  Z(Ket-address, R) ← Z(ket-address, R) +
                      O(rs)*Q(rs, C/C', M/M')*T(Bra-address, R)
                Loop end
              Loop end
            Loop end
          Loop end
        Loop end
      Loop end
    Loop end
  Loop end
Loop end
```

Fig. 1. Schematic loop layers for the integral-driven sigma vector construction. Addresses of the bra and ket are determined through the respective "function" of the variables listed in the figure. Two vector-additions are carried out according to the calculated addresses. $\boldsymbol{O}$ and $\boldsymbol{Q}$ respectively denote the arrays of integrals and coupling coefficients for the given pq

integrals (regardless of their actual existence on a file), as discussed by Siegbahn for the MRSCI case [16]. Note that it may also be possible to neglect certain costly CSF interactions. These techniques can be effective in CI calculations for large molecular systems.

The CPU cost for the OEI processings is expected to be very cheap. The structure of loops is essentially the same (except for the missing pq) as Fig. 1 for TEI but is much simpler because of the two-indexed nature. Recall that the total number of loops is only 144. The contributions from OEIs may be accumulated to the sigma vectors after the TEI processings are finished.

Once the sigma vector set $\boldsymbol{Z}$ is constructed, the energies are evaluated and the convergence is checked [17, 18]. If the procedure has not yet converged, the $\boldsymbol{Z}$ construction is iterated with the new $\boldsymbol{T}$ updated by proper corrections. After the convergence is found, the density matrix generation by using the symbolic expressions for OEIs is an easy task. The size-consistency would be a crucial requirement even for medium-sized molecules [19]. The size-consistently modified CIs like the coupled pair functional (CPF) family [20, 21] can also be calculated from $\boldsymbol{Z}$, and the process is iterated similarly until convergence.

2.6 Parallel applicability

Finally, from a viewpoint of the recent parallel computing movement (for example, see Ref. [22]), the parallel applicability of our CI formulation is discussed. In fact, the parallelism has been one of the hot topics in the 1990's MO calculations. The special issue of Theor. Chim. Acta, Vol. 84, in 1993, has nicely compiled the fruitful results of extensive efforts [23]. The parallelization of the present integral-driven sigma vector construction can be straightforwardly done by the outermost pq or, more simply, only p in Fig. 1. Note here that the CI parallelization reported by Schüler et al. in Ref. [23] was done not in the integral-driven context but in the "segmental-vector-driven" context.

As in the case of parallelized self-consistent-field (SCF) calculations in which the Fock matrix elements are constructed in parallel (refer to Ref. [23]), the parallelized sigma vector construction can obey the so-called client/server paradigm with the message-passing tools such as the TCGMSG [24] and PVM [25], where each node (or parallel processor element) has the local memory (LM). The integral file may be presegmented or properly replicated among the nodes before the iteration procedure starts. The symbolic expressions whose size is minimized may also be replicated. In a certain iteration, first the $\boldsymbol{T}$ should be broadcast by the client to the servers that are used in parallel, and the working $\boldsymbol{Z}$ area should be cleared by each server process. Second, the client assigns the "partial" index list of integrals $\{pq\}$ or $\{p\}$ to the servers. Then, each of the servers constructs "partially" the $\boldsymbol{Z}$ elements on the LM, according to the respectively assigned list. After the parallel processing is completed, the client sums up the "partial" elements of servers to reproduce the "full" $\boldsymbol{Z}$ as

$$\boldsymbol{Z}(\text{"full"})_{\text{client}} = \sum_{\text{server}} \boldsymbol{Z}(\text{"partial"})_{\text{server}} \tag{18}$$

and the CI energies are evaluated by the client. The above scheme of parallelization can lead to an efficient parallel acceleration, because the processings in the server are expected to be complicated and time-consuming due to the many layers of loops as shown in Fig. 1 and, furthermore, any intercommunication between servers is unnecessary. Each server node executes asynchronously the different loops of integral processings [Eqs. (8) and (9)]. Thus, the present CI parallelism is characterized by the "multi-instruction multi-data" (MIMD) [22].

Preliminary tests of integral-driven parallel calculations on the clustered WSs will be described in Appendix A. The acceleration efficiency shown in the appendix indicates that the integral-driven parallelism indeed is a promising recipe. We would expect that the parallel CI methodology makes the "currently inaccessible problems" of large molecules having more than a few hundred correlating MOs accessible on the forthcoming supercomputers of the vector-parallel type, where the "on-the-fly" integral usage [26] (which will be outlined in Appendix B) is probably incorporated.

3 Conclusion

In this paper, a new formulation of the integral-driven direct CI using the internally and externally symbolic energy expressions was proposed. Multi-indexed quantities like molecular integrals are systematically classified. The resulting structure of

expressions is fairly complicated. The number of unique loops for two-electron integral processings in the sigma vector construction is as many as 1325. The present formulation is oriented toward the investigations of large molecular systems by using flexibly defined CI wavefunctions. The parallel recipe is straightforward. Works of the system implementation are in progress.

Acknowledgements. We thank Drs. N. Koike, R. Nakazaki, and T. Nakata for fruitful discussions, especially on parallelism. Thanks are also due to Drs. H. Rangu, M. Yamamoto, H. Igarashi, and Y. Ohta for support. Finally, YM is thankful to Prof. K. Tanaka (University of Electro-Communications) for hearty encouragement in the early stages of this study.

Appendix A

The potential of the integral-driven parallel sigma vector construction is checked on the ethernet-clustered WSs. For this purpose, we wrote a prototype of the closed-shell single-reference SDCI program in FORTRAN (about 10000 lines), where no special code tuning was made. Note that the loop layer of C/C' is apparently unnecessary. In this prototype code, the number of unique TEI loop types to be coded is reduced to 355, because there are only five symbolic excited CSF types (4 and 13–16) and four integral types (11–15). All the energy expressions are installed in the program. The parallelization is done in the simplest round-robin fashion [23] and according to only p, which specifies both the internal (type 11 integrals) and external (type 12–15) indices at the first position of the quartet. The PVM message-passing [25] is used for the parallel control. Before the iteration, the TEI list is broadcast by the client and is kept in the LM associated with each server node.

The test molecule was the simplest amino acid, glycine (H_2N–CH_2–COOH). A standard double-zeta (DZ) basis set given by van Duijneveldt [27] was used, where the contractions are $(9s5p)/[4s2p]$ for the C, N, and O atoms and $(5s)/[2s]$ for the H atom. The molecular coordinates were taken from the experimental data [28]. The total number of AOs was 60 and the SCF energy was -282.67881 a.u. The numbers of internal and external MOs were respectively 15 and 40 and the expansion length of CI was 180901. The convergence was obtained after seven iterations of the sigma vector construction and the resulting correlation energy was -0.45633 a.u.

Table 12. Timings [s] for glycine SDCI calculations on a homogeneously clustered workstations

No. of nodes	1	2	3	4	5	6	7	8
CPU time	6288							
Elapsed time	6351	3329	2480	1850	1550	1408	1387	1130
Acceleration[b]		1.89	2.54	3.40	4.06	4.47	4.53	5.56
CPU time sum for $\boldsymbol{Z}$	6177							
Elapsed time for $\boldsymbol{Z}$	6240	3095	2331	1701	1392	1248	1217	945
Acceleration		2.00	2.65	3.63	4.44	4.95	5.08	6.54

[a] Each node is NEC EWS-4800 (360AD) system, whose MIPS and SPECfp92 values are respectively 149 and 94

[b] Defined as the ratio of the CPU time for single-node execution to the elapsed time for parallel execution

Table 12 shows the effect of parallelization in homogeneously clustered WSs, where this platform was dedicated only to the test jobs (or in dedicated usage). The acceleration for the sigma vector construction was 2.00 (1.89) for two nodes, 3.63 (3.40) for four, 4.95 (4.47) for six, and 6.54 (5.56) for eight, where the value in parentheses is for the total (or client) job. We also checked the case of heterogeneously clustered platforms. For five nodes in which there was a speed difference of more than two and was in non-dedicated usage, a parallel acceleration of about four relative to the non-parallel execution on the slowest node was observed for the total job. The overall parallel efficiency is satisfactory enough to demonstrate the potential of the integral-driven CI parallelism, especially when considering that the prototype program code was not optimum and its parallel control was somewhat naive.

Appendix B

In this paper, it has been postulated that the prescreened TEI list can be kept on a file. However, the "on-the-fly" integral usage should be desired for huge scale problems having more than a few hundred correlating orbitals. The "on-the-fly" method was pioneered by Almlöf in the SCF calculation [26]. The "on-the-fly" SCF algorithm dramatically broke the storage and retrieval bottleneck of AO integrals at the cost of repeated integral generations, and the parallel SCF calculations [23] have adopted this algorithm. Note that Taylor analyzed generally the "on-the-fly" way of TEI (MO-based integral) generation for the post-SCF or correlated calculations [29]. This appendix shows that the "on-the-fly" algorithm is naturally incorporated into our integral-driven sigma vector construction scheme with the parallelism.

Consider first the non-parallel or single-node case. Figure 2 schematically illustrates the loop structure for "on-the-fly" integral processing. The groupings seen in the figure may be adjusted according to the available memory size of the node. The outermost loop is driven by the grouped q, denoted q_g. The "not-full" AO integral generation is carried out under the control of μ_g, ν_g. λ, and σ. From the given chunk of AO integrals, the 2/4 partial transformation

$$g_{\mu_g \nu_g, rs} = \sum_{\lambda\sigma} c_{\lambda r} c_{\sigma s} g_{\mu_g \nu_g, \lambda\sigma} \tag{19}$$

is performed, where the c is the AO–MO coefficients matrix. When the loop over ν_g is finished, the list of $g_{\mu_g q_g, rs}$ is available. The rest of the transformation is for p. Once the proper prescreening is done, the inner process is virtually the same as shown in Fig. 1, except that the completely transformed integral $g_{pq,rs}$ is replaced by the product of the AO–MO coefficient and the 3/4 transformed integral $c_{\mu p} g_{\mu q, rs}$. Then the loop control goes to the next μ_g. The cost to be paid is manifestly that "full" AO integral generation is repeated as many times as there are q groups. Furthermore, the CPU time per one "full' calculation can now be increased fourfold because the eightfold equivalent permutations of AO integral indices are degraded by the fixing of μ and ν groupings to only twofold permutations.

Parallelization of the integral-driven/"on-the-fly" sigma vector construction is obvious. The q_g that drives the outermost loop in Fig. 2 is taken as the parallel parameter. Good parallel efficiency would be expected.

Loop over q_g
  Loop over μ_g
    Loop over ν_g
      Loop over $\lambda\sigma$
        Calculate AO integrals of $g_{\mu_g\nu_g,\lambda\sigma}$
      Loop end
      Loop over rs
        Perform 2/4 transformation of $g_{\mu_g\nu_g,rs} = \sum_{\lambda\sigma} C_{\lambda r} C_{\sigma s} g_{\mu_g\nu_g,\lambda\sigma}$
    Loop end
    Loop over $\nu \in \nu_g$
      Perform accumulation of $g_{\mu_g q_g,rs} \leftarrow g_{\mu_g q_g,rs} + c_{\nu q_g} g_{\mu_g \nu,rs}$
    Loop end
  Loop end
  Loop over $q \in q_g$
    Loop over $\mu \in \mu_g$
      Loop over p
        Calculate contributions to sigma vectors from $c_{\mu p} g_{\mu q,rs}$
      Loop end
    Loop end
  Loop end
Loop end

Fig. 2. Schematic loop layers for "on-the-fly" integral usage. Inside from the p loop, processing is virtually the same as that indicated in Fig. 1 (needless to say, integral reading is unnecessary)

References

1. Roos B (1972) Chem Phys Lett 15:153
2. Roos BO, Siegbahn PEM (1977) In: Shaefer III HF (ed) Modern theoretical chemistry. Plenum, New York
3. Hinze J (ed) (1981) Lecture Notes in Chemistry 22, The Unitary Group. Springer, Berlin
4. Dykstra CE (ed) (1984) Advanced theories and computational approaches to the electronic structure of molecules. Reidel, Dordrecht
5. Lawley KP (ed) (1987) Advances in Chemical Physics 59, *Ab initio* methods in quantum chemistry, Part 2, Wiley, Chichester
6. Roos BO (ed) (1992) Lecture Notes in Chemistry 58, Europe Summer School in quantum chemistry, Springer, Berlin
7. Liu B, Yoshimine M (1981) J Chem Phys 74:612
8. Saunders VR, van Lenthe JH (1983) Mol Phys 48:923
9. Pauncz R (1979) Spin eigenfunctions. Plenum, New York
10. McLean AD, Yoshimine M (1973) J Chem Phys 58:1066
11. Tatewaki H, Tanaka K, Sasaki F, Obara S, Ohno K, Yoshimine M (1979) Int J Quant Chem 15:533
12. Wilson S (1984) Electron correlation in molecules. Clarendon, Oxford
13. Lindgren I, Morrison J (1986) Atomic many body theory, 2nd ed. Springer, Berlin
14. Tanaka K (1972) Electronic structure of the ground state, low-lying excited (valence and Rydberg type) states, and the first ionized state of the formaldehyde PhD thesis, Faculty of Science, Hokkaido University, Sapporo
15. Weinstein H, Pauncz R, Cohen M (1971) Adv Atom Mol Phys 7:97
16. Siegbahn PEM (1979) J Chem Phys 70:391
17. Davidson ER (1975) J Comp Phys 17:87
18. Kosugi N (1984) J Comp Phys 55:426
19. Szabo A, Ostlund NS (1982) Modern quantum chemistry. Macmillan, New York
20. Ahlrichs R, Scharf P, Ehrhardt C (1985) J Chem Phys 82:890

21. Gdanitz RJ, Ahlrichs R (1988) Chem Phys Lett 143:413
22. Hord RM (1993) Parallel supercomputing in MIMD architectures. CRC, Boca Raton
23. Theor Chim Acta 84 (1993) a special issue for parallel MO calculations. Technical terms for the parallel computings (such as the "client/server") can be found in this issue
24. TCGMSG (theoretical chemistry group's message-passings): a portable message-passing toolkit, Horrison RJ (1991) Int J Quant Chem 40:847
25. PVM (parallel virtual machine): a message-passing system used most widely in the world, Beguelin ALK, Dongarra JJ, Geist GA, Jiang WC, Manchek RJ, Moore BK, Sunderam VS (1993) PVM 3.3 User's Guide and Reference Manual, available through the world-wide web (WWW). The address is http://www.epm.ornl.gov/pvm/, Oak Ridge National Laboratory, USA
26. Almlöf J, Faegri Jr K, Korsell K (1982) J Comp Chem 3:385. They originally used the name "direct SCF". If one considers the direct-CI, which calculates the integrals in every iteration, the name would correspond to "direct–direct CI", which, as noted by Taylor [29], is not a good name. Thus, in the present text, the term "on-the-fly" is used for the direct integral usage
27. van Duijneveldt FB (1971) IBM Research Laboratory Report RJ945, IBM Corp, San Jose
28. Coordinates are taken from the internal database of the BIOCESS[E] computer-aided protein/drug design system. NEC Corp, Tokyo
29. Taylor PR (1987) Int J Quant Chem 31:521